# PART III. DETECTION MONITORING TESTS

This third part of the Unified Guidance presents core procedures recommended for formal detection monitoring at RCRA-regulated facilities. **Chapter 16** describes two-sample tests appropriate for some small facilities, facilities in interim status, or for periodic updating of background data. These tests include two varieties of the *t*-test and two non-parametric versions-- the Wilcoxon rank-sum and Tarone-Ware procedures. **Chapter 17** discusses one-way analysis of variance [ANOVA], tolerance limits, and the application of trend tests during detection monitoring. **Chapter 18** is a primer on several kinds of prediction limits, which are combined with retesting strategies in **Chapter 19** to address the statistical necessity of performing multiple comparisons during RCRA statistical evaluations. Retesting is also discussed in **Chapter 20**, which presents control charts as an alternative to prediction limits.

As discussed in **Section 7.5**, any of these detection-level tests may also be applied to compliance/assessment and corrective action monitoring, where a background groundwater protection standard [GWPS] is defined as a critical limit using *two- or multiple-sample* comparison tests. Caveats and limitations discussed for detection monitoring tests are also relevant to this situation. To maintain continuity of presentation, this additional application is presumed but not repeated in the following specific test and procedure discussions.

Although other users and programs may find these statistical tests of benefit due to their wider applicability to other environmental media and types of data, the methods described in **Parts III** and **IV** are primarily tailored to the RCRA setting and designed to address formal RCRA monitoring requirements. In particular, the series of prediction limit tests found in **Chapter 18** is designed to address the range of interpretations of the sampling rules in §264.97(g), §264.98(d) and §258.54. Further, *all* of the regulatory tests listed in §264.97(i) and §258.53(h) are discussed, as well as the Student's *t*-test requirements of §265.93(b).

Taken as a whole, the set of detection monitoring methods presented in the Unified Guidance should be appropriate for almost all the situations likely to be encountered in practice. Professional statistical consultation is recommended for the rest.

*This page intentionally left blank*

# CHAPTER 16. TWO-SAMPLE TESTS



This chapter describes statistical tests between two groups of data, known as two-sample tests. These tests may be appropriate for the smallest of RCRA sites performing upgradient-to-downgradient comparisons on a very limited number of wells and constituents. They may also be required for certain facilities in interim status, and can be more generally used to compare older versus newer data when updating background.

Two versions of the classic Student's *t*-test are first discussed: the ***pooled variance t-test*** and ***Welch's t-test.*** Since both these tests expect approximately normally-distributed data as input, two non-parametric alternatives to the *t*-test are also described: the ***Wilcoxon rank-sum test*** (also known as the ***Mann-Whitney***) and the ***Tarone-Ware test.*** The latter is particularly helpful when the sample data exhibit a moderate to larger fraction of non-detects and/or multiple detection/reporting limits.

## 16.1 PARAMETRIC T-TESTS

### BACKGROUND AND PURPOSE

A statistical comparison between two sets of data is known as a two-sample test. While several varieties of two-sample tests exist, the most common is the parametric *t*-test. This test compares two distinct statistical populations. The goal of the two-sample *t*-test is to determine whether there is any statistically significant difference between the mean of the first population when compared against the mean of the second population, based on the results observed in the two respective ***samples***.

In groundwater monitoring, the typical hypothesis at issue is whether the average concentration at a compliance point is the same as (or less than) the average concentration in background, or whether the compliance point mean is larger than the background mean, as represented in equation [16.1] below:

$$H_0 : \mu_C \le \mu_{BG} \quad vs. \quad H_A : \mu_C > \mu_{BG} \qquad [16.1]$$

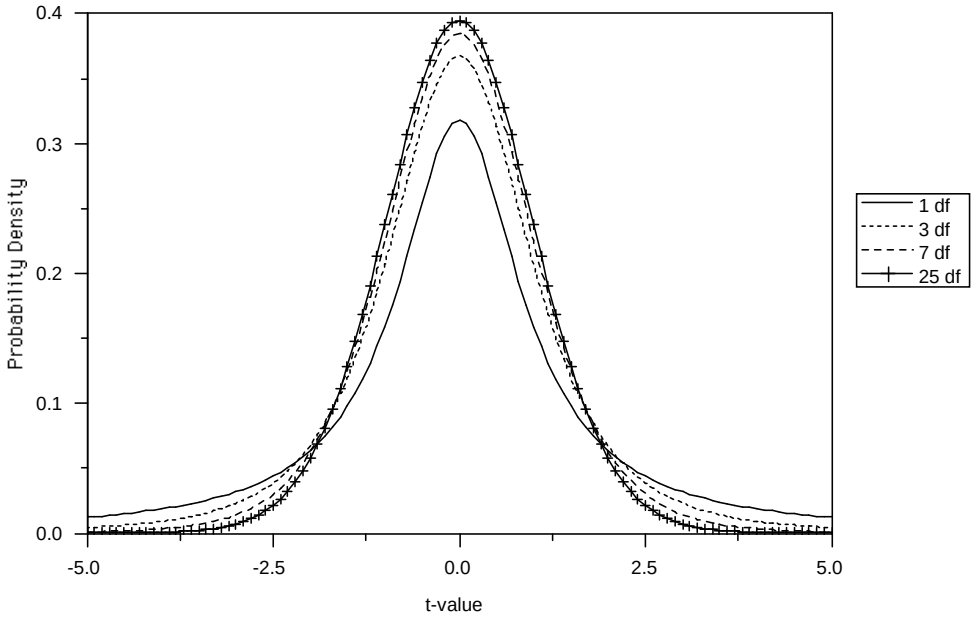
A natural statistic for comparing two population means is the difference between the sample means, $\left(\bar{x}_C - \bar{x}_{BG}\right)$. When this difference is small, a real difference between the respective population means is considered unlikely. However, when the sample mean difference is large, the null hypothesis is rejected, since in that case a real difference between the populations seems plausible. Note that an observed difference between the ***sample*** means does ***not*** automatically imply a true population difference. Sample means can vary for many reasons even if the two underlying parent populations are

identical. Indeed, the Student's *t*-test was invented precisely to determine when an observed sample difference should be considered significant (*i.e.*, more than a chance fluctuation), especially when the sizes of the two samples tend to be small, as is the usual case in groundwater monitoring.

Although the null hypothesis ($H_0$) represented in equation [16.1] allows for a true compliance point mean to be less than background, the behavior of the *t*-test statistic is assessed at the point where $H_0$ is most difficult to verify — that is, when $H_0$ is true and the two population means are identical. Under the assumption of equal population means, the test statistic in any *t*-test will tend to follow a Student's *t*-distribution. This fact allows the selection of critical points for the *t*-test based on a pre-specified Type I error or false positive rate ($\alpha$). Unlike the similarly symmetric normal distribution, however, the Student's *t*-distribution also depends on the number of independent sample values used in the test, represented by the ***degrees of freedom*** [*df*].

The number of degrees of freedom impacts the shape of the *t*-distribution, and consequently the magnitude of the critical (percentage) points selected from the *t*-distribution to provide a basis of comparison against the *t*-statistic (see **Figure 16-1**). In general, the larger the sample sizes of the two groups being compared, the larger the corresponding degrees of freedom, and the smaller the critical points (in absolute value) drawn from the Student's *t*-distribution. In a one-sided hypothesis test of whether compliance point concentrations exceed background concentrations, a smaller critical point corresponds to a more powerful test. Therefore, all other things being equal, the larger the sample sizes used in the two-sample *t*-test, the more protective the test will be of human health and the environment.

Figure 16-1. Student's *t*-Distribution for Varying Degrees of Freedom




In groundwater monitoring, *t*-tests can be useful in at least two ways. First, a *t*-test can be employed to compare background data from one or more upgradient wells against a single compliance

well. If more than one background well is involved, all the upgradient data would be pooled into a single group or sample before applying the test.

Second, a *t*-test can be used to assess whether updating of background data is appropriate (see **Chapter 5** for further discussion). Specifically, the two-sample *t*-test can be utilized to check whether the more recently collected data is consistent with the earlier data assigned initially as the background data pool. If the *t*-test is non-significant, both the initial background and more recent observations may be considered part of the same statistical population, allowing the overall background data set to grow and to provide more accurate information about the characteristics of the background population.

The Unified Guidance describes two versions of the parametric *t*-test, the pooled variance Student's *t*-test and a modification to the Student's *t*-test known as Welch's *t*-test. This guidance prefers the latter *t*-test to use of Cochran's Approximation to the Behrens-Fisher (CABF) Student's *t*-test. Initially codified in the 1982 RCRA regulations, the CABF *t*-test is no longer explicitly cited in the 1988 revision to those regulations. Both the pooled variance and Welch's *t*-tests are more standard in statistical usage than the CABF *t*-test. When the parametric assumptions of the two-sample *t*-test are violated, the Wilcoxon rank-sum or the Tarone-Ware tests are recommended as non-parametric alternatives.

## REQUIREMENTS AND ASSUMPTIONS

The two-sample *t*-test has been widely used and carefully studied as a statistical procedure. Correct application of the Student's *t*-test depends on certain key assumptions. First, every *t*-test assumes that the observations in each data set or group are statistically independent. This assumption can be difficult to check in practice (see **Chapter 14** for further discussion of statistical independence), especially if only a handful of measurements are available for testing. As noted in **Chapter 5** in discussing data mixtures, lab replicates or field duplicates are not statistically independent and should not be treated as independent water quality samples. That section discussed the limited conditions under which certain replicate data might be applicable for *t*- testing. Incorrect usage of replicate data was one of the concerns that arose in the application of the CABF *t*-test.

Second, all *t*-tests assume that the underlying data are approximately normal in distribution. Checks of this assumption can be made using one of the tests of normality described in **Chapter 10**. The *t*-test is a reasonably robust statistical procedure, meaning that it will usually provide accurate results even if the assumption of normality is partially violated. This robustness of the *t*-test provides some insurance against incorrect test results if the underlying populations are non-normal. However, the robust assumption is dubious when the parent population is heavily skewed. For data that are lognormal and positively skewed, the two-sample *t*-test can give misleading results unless the data are first log-transformed. Similarly, a transformation may be needed to first normalize data from other non-normal distributions.

Another assumption particularly relevant to the use of *t*-tests in groundwater monitoring is that the population means need to be *stable* or *stationary* over the time of data collection and testing. As discussed in **Part II** of the guidance, many commonly monitored groundwater parameters exhibit mean changes in both space and time. Consequently, correct application of the *t*-test in groundwater requires an implicit assumption that the two populations being sampled (*e.g.*, a background well and a

compliance point well) have average concentrations that are not trending with time. Time series plots and diagnostic trend tests (**Chapter 14**) can sometimes be used to check this assumption.

The *t*-test does an excellent job of identifying a stable mean level difference between two populations. However, if one or both populations have trends observable in the sample measurements, the *t*-test may have difficulty correctly identifying a difference between the two groups. For instance, if earlier samples in a compliance well were uncontaminated but later samples are increasing with time, the *t*-test may still provide a non-significant result. With compliance point concentrations increasing relative to background, the *t*-test may not be the appropriate method for identifying this change. Some form of trend testing will provide a better evaluation.

Another concern in applying the *t*-test to upgradient-downgradient interwell comparisons is that the null hypothesis is assumed to be true ***unless*** the downgradient well becomes contaminated. Absent such an impact, the population means are implicitly assumed to be identical. ***Spatial variability*** in background and compliance well groundwater concentrations for certain monitoring constituents do not allow clear conditions for comparisons intended to identify a release at a downgradient compliance well. Natural or pre-existing synthetic mean differences among background wells will be confused with a potential release. In such cases, neither the two-sample *t*-test nor ***any*** interwell procedure comparing upgradient against downgradient measurements is likely to give a correct conclusion.

One final requirement for running any *t*-test is that each group should have an adequate sample size. The *t*-test will have minimal statistical power to identify any but the largest of concentration differences if the sample size in each group is less than four. Four measurements per group should be considered a ***minimum*** requirement, and much greater power will accrue from larger sample sizes. Of course, the attractiveness of larger data sets must be weighed against the need to have statistically independent samples and the practical limitation of semi-annual or annual statistical evaluations. These latter requirements often constrain the frequency of sampling so that it may be impractical to secure more than 4 to 6 or possibly 8 samples during any annual period.

## 16.1.1 POOLED VARIANCE T-TEST

### BACKGROUND AND PURPOSE

In the case of two independent samples from normal populations with common variance, the Student's *t*-test statistic is expressed by the following equation:

$$t = \left(\bar{x}_C - \bar{x}_{BG}\right) \Bigg/ \sqrt{\left[\frac{\left(n_{BG}-1\right)s_{BG}^2 + \left(n_C-1\right)s_C^2}{\left(n_{BG}+n_C-2\right)}\right]\left(\frac{1}{n_{BG}}+\frac{1}{n_C}\right)} \qquad [16.2]$$

The first bracketed quantity in the denominator is known as the ***pooled variance***, a weighted average of the two sample variances. The entire denominator of equation [16.2] is labeled the ***standard error of the difference*** ($SE_{diff}$). It represents the probable chance fluctuation likely to be observed between the background and compliance point sample means when the null hypothesis in equation [16.1] is true. Note that the formula for $SE_{diff}$ depends on both the pooled variance and the sample size of each group.

When the null hypothesis ($H_0$) is satisfied and the two populations are truly identical, the test statistic in equation [16.2] behaves according to an exact Student's *t*-distribution. This fact enables critical points for the *t*-test to be selected based on a pre-specified Type I error rate (α) and an appropriate degrees of freedom. In equation [16.2], the joint degrees of freedom is equal to $\left(n_{BG} + n_C - 2\right)$, the sum of the background and compliance point sample sizes less two degrees of freedom (one for each mean estimate).

## REQUIREMENTS AND ASSUMPTIONS

Along with the general requirements for *t*-tests, the pooled variance version of the test assumes that the population variances are equal in both groups. Since only the sample variances will be known, this assumption requires a formal statistical test of its own such as Levene's test described in **Chapter 11**. An easier, descriptive method is to construct side-by-side box plots of both data sets. If the population variances are equal, the interquartile ranges represented by the box lengths should also be comparable. If the population variances are distinctly different, on the other hand, the box lengths should also tend to be different, with one box much shorter than the other.

When variances are unequal, the Unified Guidance recommends Welch's *t*-test be run instead. Welch's *t*-test does not require the assumption of equal variances across population groups. Furthermore, the performance of Welch's *t*-test is almost always equal or superior to that of the usual Student's *t*-test. Therefore, one may be able to skip the test of equal variances altogether before running Welch's *t*-test.

All t-tests require approximately normally-distributed data. If a common variance ($\sigma^2$) exists between the background and compliance point data sets, normality in the pooled variance *t*-test can be assessed by examining the combined set of background and compliance point ***residuals.*** A residual can be defined as the difference between any individual value and its sample group mean (***e.g.***, $x_i - \bar{x}_{BG}$ for background values $x_i$). Not only will the combined set of residuals allow for a more powerful test of normality than if the two samples are checked separately, but it also avoids a difficulty that can occur if the sample measurements are naively evaluated with the ***Shapiro-Wilk multiple group test***. The multiple group normality test allows for populations with different means ***and*** different variances. If an equal variance check has not already been made, the multiple group test could register both populations as being normal even though the two population variances are distinctly different. The latter would violate a key assumption of the pooled variance t-test. To avoid this potential problem, either always check explicitly for equal variances before running the pooled variance *t*-test, or consider running Welch's *t*-test instead.

## PROCEDURE

Step 1. To conduct the two-sample Student's *t*-test at an α-level of significance, first compute the sample mean ($\bar{x}$) and standard deviation (*s*) of each group. Check for equal variances using a test from **Chapter 11**. If there is no evidence of heteroscedasticity, check normality in both samples, perhaps by calculating the residuals from each group and running a normality test on the combined data set.

Step 2. Once the key assumptions have been checked, calculate the two-sample *t*-statistic in equation [16.2], making use of the sample mean, sample standard deviation, and sample size of each group.

Step 3. Set the degrees of freedom to $df = n_{BG} + n_C - 2$, and look up the $(1-\alpha) \times 100$th percentage point from the *t*-distribution in **Table 16-1** in **Appendix D**. Compare this $\alpha$-level critical point against the *t*-statistic. If the *t*-statistic does not exceed the critical point, conclude there is insufficient evidence of a significant difference between the two population means. If, however, the *t*-statistic is greater than the critical point, conclude that the compliance point population mean is significantly greater than the background mean.

### ►EXAMPLE 16-1

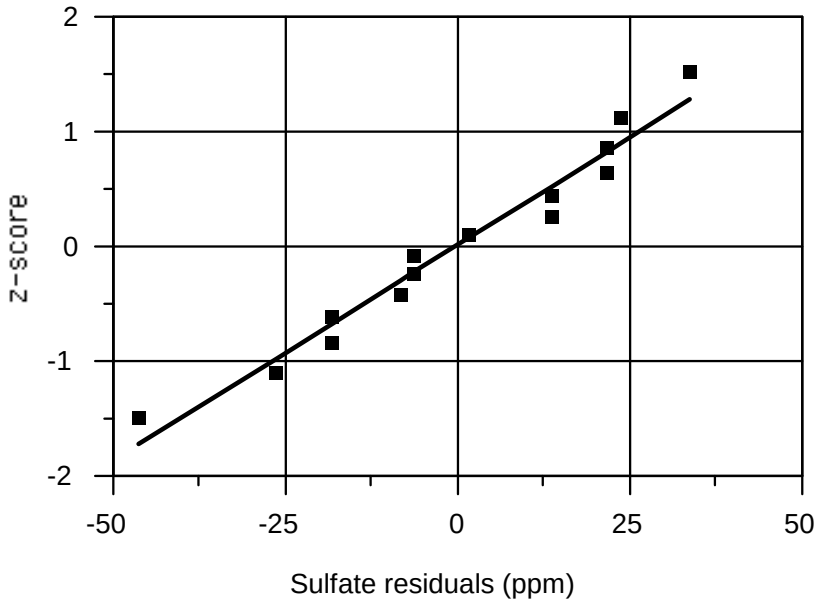
Consider the quarterly sulfate data in the table below collected from one upgradient and one downgradient well during 1995-96. Use the Student's *t*-test to determine if the downgradient sulfate measurements are significantly higher than the background values at an $\alpha = 0.01$ significance level.

| | **Sulfate Concentrations (ppm)** | | | |
|---|---|---|---|---|
| Quarter | Background | Downgradient | Background Residuals | Downgradient Residuals |
| 1/95 | 560 | | 23.75 | |
| 4/95 | 530 | | −6.25 | |
| 7/95 | 570 | 600 | 33.75 | −8.33 |
| 10/95 | 490 | 590 | −46.25 | −18.33 |
| 1/96 | 510 | 590 | −26.25 | −18.33 |
| 4/96 | 550 | 630 | 13.75 | 21.67 |
| 7/96 | 550 | 610 | 13.75 | 1.67 |
| 10/96 | 530 | 630 | −6.25 | 21.67 |
| Mean | 536.25 | 608.33 | | |
| SD | 26.6927 | 18.3485 | | |

SOLUTION

Step 1. Compute the sample mean and standard deviation in each well, as listed in the table above. Then compute the sulfate residuals by subtracting the well mean from each individual value. These differences are also listed above. Comparison of the sample variances shows no evidence that the population variances are unequal. Further, a probability plot of the combined set of residuals (**Figure 16-2**) indicates that the normal distribution appears to provide a reasonable fit to these data.

Figure 16-2. Probability Plot of Combined Sulfate Residuals




Step 2. Compute the two-sample *t*-statistic on the raw sulfate measurements using equation [16.2]. Note that the background sample size is $n_{BG}$ = 8 and the downgradient sample size is $n_C$ = 6.

$$t = \left(608.33 - 536.25\right) \Bigg/ \sqrt{\left[\frac{7\left(26.6927\right)^2 + 5\left(18.3485\right)^2}{8+6-2}\right]\left(\frac{1}{8}+\frac{1}{6}\right)} = 5.66$$

Step 3. Compute the degrees of freedom as $df = 8 + 6 - 2 = 12$. Since $\alpha = .01$, the critical point for the test is the upper 99th percentile of the *t*-distribution with 12 *df*. **Table 16-1** in **Appendix D** then gives the value for $t_{cp} = 2.681$. Since the *t*-statistic is clearly larger than the critical point, conclude the downgradient sulfate population mean is significantly larger than the background population mean at the 0.01 level. ◄

## 16.1.2 WELCH'S T-TEST

### BACKGROUND AND PURPOSE

The pooled variance Student's *t*-test in **Section 16.1.1** makes the explicit assumption that both populations have a common variance, $\sigma^2$. For many wells and monitoring constituents, local geochemical conditions can result in both different well means and variances. A contamination pattern at a compliance well can have very different variability than its background counterpart.

Welch's *t*-test was designed as a modification to the Student's *t*-test when the population variances might differ between the two groups. The Welch's *t*-test statistic is defined by the following equation:

$$t = \left(\bar{x}_C - \bar{x}_{BG}\right) \Big/ \sqrt{\frac{s_{BG}^2}{n_{BG}} + \frac{s_C^2}{n_C}} \qquad [16.3]$$

The denominator of equation [16.3] is also called the ***standard error of the difference*** ($SE_{\text{diff}}$), similar to the pooled variance *t*-test. But it is a different weighted estimate based on the respective sample variances and sample sizes, reflecting the fact that the two population variances may not be the same.

The most difficult part of Welch's *t*-test is deriving the correct degrees of freedom. Under the assumption of a common variance, the pooled variance estimate incorporated into the usual Student's *t*-test has $df = \left(n_{BG} + n_C - 2\right)$ degrees of freedom, representing the number of independent "bits" of sample information included in the variance estimate. In Welch's *t*-test, the derivation of the degrees of freedom is more complicated, but can be approximately computed with the following equation:

$$\hat{df} = \left[\frac{s_{BG}^2}{n_{BG}} + \frac{s_C^2}{n_C}\right]^2 \Big/ \left[\frac{\left(s_{BG}^2/n_{BG}\right)^2}{n_{BG}-1} + \frac{\left(s_C^2/n_C\right)^2}{n_C-1}\right] \qquad [16.4]$$

Despite its lengthier calculations, Welch's *t*-test has several practical advantages. Best and Rayner (1987) found that among statistical tests specifically designed to compare two populations with different variances, Welch's *t*-test exhibited comparable statistical power (for $df \geq 5$) and was much easier to implement in practice than other tests they examined. Moser and Stevens (1992) compared Welch's *t*-test against the usual pooled variance *t*-test and determined that Welch's procedure was the more appropriate in almost every case. The only advantage registered by the usual Student's *t*-test in their study was in the case where the sample sizes in the two groups were unequal ***and*** the population variances were ***known*** to be essentially the same. In practice, the population variances will almost never be known in advance, so it appears reasonable to use Welch's *t*-test in the majority of cases where a two-sample *t*-test is warranted.

## REQUIREMENTS AND ASSUMPTIONS

Welch's *t*-test is also a reasonably robust statistical procedure, and will usually provide accurate results even if the assumption of normality is partially violated. This robustness of the *t*-test provides some insurance against incorrect test results if the underlying populations are non-normal. But heavily skewed distributions do require normalizing transformations. Certain limitations apply when using transformed data, discussed in the following section.

Unlike the pooled variance *t*-test, Welch's procedure does not require that the population variances be equal in both groups. Other general requirements of *t*-tests, however, such as statistical independence of the sample data, lack of spatial variability when conducting an interwell test, and stationarity over time, are applicable to Welch's t-test and needs to be checked prior to running the procedure.

Because the variances of the tested populations may not be equal, an assessment of normality cannot be made under Welch's t-test by combining the residuals (as with the pooled variance t-test), unless an explicit check for equal variances is first conducted. The reason is that the combined residuals from normal populations with different variances may not test as normal, precisely because of the

heteroscedasticity. Since this latter variance check is not required for Welch's test, it may be easier to input the sample data directly into the *multiple group test of normality* described in **Chapter 10**.

## PROCEDURE

Step 1. To run the two-sample Welch's *t*-test, first compute the sample mean $(\bar{x})$, standard deviation ($s$), and variance ($s^2$) in each of the background ($BG$) and compliance point ($C$) data sets.

Step 2. Compute Welch's *t*-statistic with equation [16.3].

Step 3. Compute the approximate degrees of freedom in equation [16.4] using the sample variance and sample size from each group. Since this quantity often results in a fractional amount, round the approximate $\hat{df}$ to the nearest integer.

Step 4. Depending on the $\alpha$ significance level of the test, look up an appropriate critical point ($t_{cp}$) in **Table 16-1** in **Appendix D**. This entails finding the upper $(1-\alpha)\times 100th$ percentage point of the Student's *t*-distribution with *df* degrees of freedom.

Step 5. Compare the *t*-statistic against the critical point. If $t \le t_{cp}$, conclude there is no statistically significant difference between the background and compliance point population means. If, however, $t > t_{cp}$, conclude that the compliance point population mean is significantly greater than the background mean at the $\alpha$ level of significance.

## ►EXAMPLE 16-2

Consider the following series of monthly benzene measurements (in ppb) collected over 8 months from one upgradient and one downgradient well. What significant difference, if any, does Welch's *t*-test find between these populations at the $\alpha = .05$ significance level?

| | **Benzene (ppb)** | |
|---|---|---|
| Month | BG | DG |
| Jan | 0.5 | 0.5 |
| Feb | 0.8 | 0.7 |
| Mar | 1.6 | 4.6 |
| Apr | 1.8 | 2.0 |
| May | 1.1 | 16.7 |
| Jun | 16.1 | 12.5 |
| Jul | 1.6 | 26.3 |
| Aug | 0.6 | 186.0 |
| N | 8 | 8 |
| Mean | 3.0 | 31.2 |
| SD | 5.31 | 63.22 |
| Variance | 28.204 | 3997.131 |

Step 1. Compute the sample mean, standard deviation, and variance of each group as in the table above.

Step 2. Use equation [16.3] to compute Welch's *t*-statistic:

$$t = \left(31.2 - 3.0\right) \Big/ \sqrt{\frac{28.204}{8} + \frac{3997.131}{8}} = 1.257$$

Step 3. Compute the approximate degrees of freedom using equation [16.4]:

$$\hat{df} = \left[\frac{28.204}{8} + \frac{3997.131}{8}\right]^2 \Big/ \left[\frac{(28.204/8)^2}{7} + \frac{(3997.131/8)^2}{7}\right] = 7.1 \quad \approx 7$$

Step 4. Using **Table 16-1** in **Appendix D** and given $\alpha$ = .05, the upper 95% critical point of the Student's *t*-distribution with 7 *df* is equal to 1.895.

Step 5. Compare the *t*-statistic against the critical point, $t_{cp}$. Since $t < t_{cp}$, the test on the raw concentrations provides insufficient evidence of a true difference in the population means. However, given the order of magnitude difference in the sample means and the fact that several of the downgradient measurements are substantially larger than almost all the background values, we might suspect that one or more of the *t*-test assumptions was violated, possibly invalidating the result. ◄

### 16.1.3 WELCH'S T-TEST AND LOGNORMAL DATA

Users should recall that if the underlying populations are *lognormal* instead of normal and Welch's *t*-test is run on the logged data, the procedure is not a comparison of arithmetic means but rather between the population *geometric means.* In the case of a lognormal distribution, the geometric means are equivalent to the population *medians*. In effect, a test of the log-means is equivalent to a test of the medians in terms of the raw concentrations. Both the population geometric mean and the lognormal median can be estimated from the logged measurements as $\exp(\bar{y})$, where $y = \log x$ represents a logged value and $\bar{y}$ is the log-mean. On the other hand, the (arithmetic) lognormal mean on the concentration scale would be estimated as $\exp\left(\bar{y} + s_y^2/2\right)$, a quantity larger than the geometric mean or median due to the presence of the term involving $s_y^2$, the log-variance.

Although a *t*-test conducted in the logarithmic domain is not a direct comparison of the arithmetic means, there are situations where that comparison can be *inferred* from the test results. For instance, consider using the pooled variance two-sample Student's *t*-test on logged data with a common (*i.e.*, equal) population log-variance ($\sigma_y^2$) in each group. In that case, finding a larger geometric mean or median in a compliance well population when compared to background also implies that the compliance point *arithmetic mean* is larger than the background *arithmetic mean*. However, when using Welch's *t*-test, the assumption of equal variances is not required. Because of this, on rare occasions one might find

a larger compliance point geometric mean or median when testing the log-transformed data, even though the compliance point population arithmetic mean is ***smaller*** than the background arithmetic mean.

Fortunately, such a reversal can only occur in the unlikely situation that the background population log-variance is distinctly larger than the compliance point log-variance. Factors contributing to an increase in the log-mean concentration level in lognormal populations often serve, if anything, to also increase the log-variance, and almost never to ***decrease*** it. Consequently, *t*-test results indicating a compliance point geometric mean higher than background should very rarely imply a less-than-background compliance point log-variance. This in turn will generally ensure that the compliance point arithmetic mean is also larger than the background arithmetic mean, so that a test of the log-transformed measurements can be used to infer whether a difference exists in the population ***concentration means***.

One caution in this discussion is for cases where the Welch's *t*-test is ***not significant*** on the log-transformed measurements. Because the log-variances ($\sigma_y^2$) are not required to be equal in the two populations when running Welch's *t*-test, yet the arithmetic lognormal mean depends on both the population log-mean ($\mu_y$) and the log-variance through the quantity $\exp\left(\mu_y + \sigma_y^2/2\right)$, it should ***not*** be inferred that a non-significant comparison on the log-scale between a compliance point and background is equivalent to finding ***no difference*** between the lognormal arithmetic ***means***. If the log-variances differ but the log-means do not, the lognormal arithmetic ***means*** will still be different even though the lognormal ***medians*** might be identical.

Therefore, if a comparison of arithmetic means is required, but the statistical populations are lognormal, care must be taken in interpreting the results of Welch's *t*-test. Two possible remedies would include: 1) only running a *t*-test on lognormal data if the log-variances can be shown to be approximately equivalent (this would allow use of the pooled variance *t*-test); and 2) using a non-parametric two-sample bootstrap procedure on the original (non-logged) measurements to compare the arithmetic means directly. Consultation with a professional statistician may be required in this second case.

►EXAMPLE 16-3

The benzene data from **Example 16-2** indicated no significant upgradient-to-downgradient difference in population means when tested on the raw measurement scale. Check to see whether the same data more closely approximate a lognormal distribution and conduct Welch's *t*-test under that assumption.

| | Benzene (ppb) | | Log(Benzene) log(ppb) | |
|---|---|---|---|---|
| Month | BG | DG | BG | DG |
| Jan | 0.5 | 0.5 | −0.693 | −0.693 |
| Feb | 0.8 | 0.7 | −0.223 | −0.357 |
| Mar | 1.6 | 4.6 | 0.470 | 1.526 |
| Apr | 1.8 | 2.0 | 0.588 | 0.693 |
| May | 1.1 | 16.7 | 0.095 | 2.815 |
| Jun | 16.1 | 12.5 | 2.779 | 2.526 |
| Jul | 1.6 | 26.3 | 0.470 | 3.270 |
| Aug | 0.6 | 186.0 | −0.511 | 5.226 |
| N | 8 | 8 | 8 | 8 |
| Mean | 3.0 | 31.2 | 0.372 | 1.876 |
| SD | 5.31 | 63.22 | 1.0825 | 1.9847 |
| Variance | 28.204 | 3997.131 | 1.1719 | 3.9392 |

SOLUTION

Step 1. First check normality of the original measurements. To do this, compute the Shapiro-Wilk statistic (*SW*) separately for each well. *SW* = 0.505 for the background data, and *SW* = 0.544 for the downgradient well. Combining these two values using the equations in **Section 10.7**, the multiple group Shapiro-Wilk statistic becomes *G* = –6.675, which is significantly less than the 5% critical point of –1.645 from the standard normal distribution.[1] Thus, the assumption of normality was violated in **Example 16-2**.

Step 2. Compute the log-mean, log-standard deviation, and log-variance of each group, as listed above. Then compute the multiple group Shapiro-Wilk test to check for (joint) normality on the log-scale. The respective *SW* statistics now increase to 0.818 for the background data and 0.964 for the downgradient well. Combining these into an overall test, the multiple group Shapiro-Wilk statistic becomes –0.721 which now exceeds the $\alpha$ = 0.05 standard normal critical point. A log transformation adequately normalizes the benzene data — suggesting that the underlying populations are lognormal in distribution — so that Welch's *t*-test can be run on the logged data.

Step 2. Using the logged measurements and equation [16.3], the *t*-statistic becomes:

$$t = \left(1.876 - 0.372\right) \Big/ \sqrt{\frac{1.1719}{8} + \frac{3.9392}{8}} = 1.88$$

[1] Note that $\alpha$ = 5% is used in this example because the total sample size (BG and DG) is *n* = 16. Nevertheless, the test would also fail at $\alpha$ = 1% or just about any significance level one might choose.

Step 3. Again using the log-variances and equation [16.4], the approximate *df* works out to:

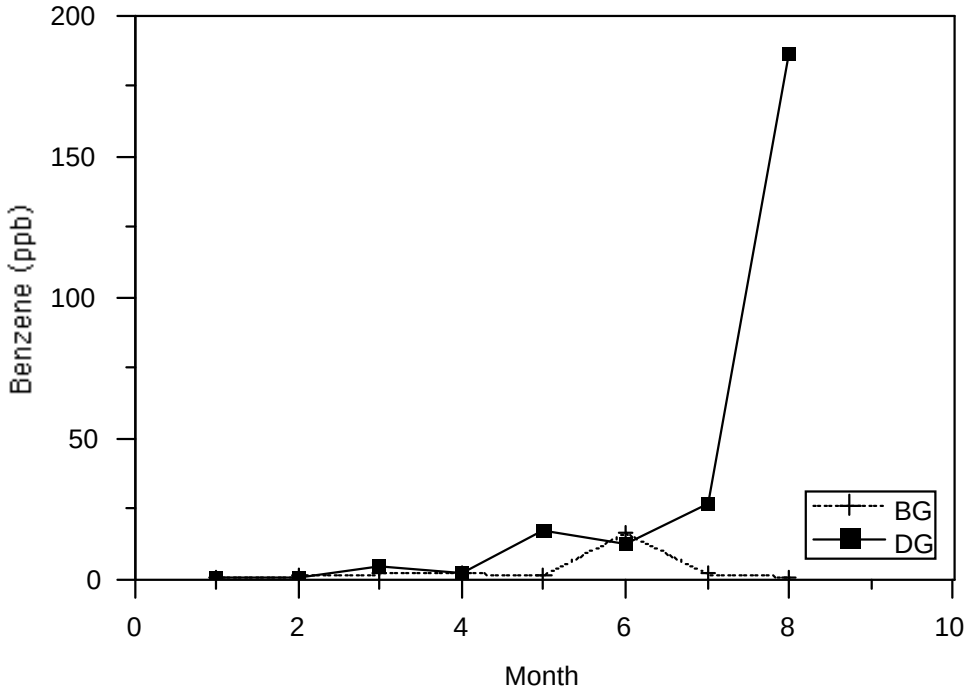
$$df = \left[\frac{1.1719}{8} + \frac{3.9392}{8}\right]^2 \Bigg/ \left[\frac{\left[1.1719/8\right]^2}{7} + \frac{\left[3.9392/8\right]^2}{7}\right] = 10.8 \approx 11$$

Note that the approximate *df* in Welch's *t*-test is somewhat less than the value that would be computed for the two-sample pooled variance Student's *t*-test. In that case, with 8 samples per data set, the *df* would have been 14 instead of 11. The reduction in degrees of freedom is due primarily to the apparent difference in variance between the two groups.

Step 4. Using **Table 16-1** in **Appendix D** and given $\alpha = .05$, the upper 95% critical point of the Student's *t*-distribution with 11 *df* is equal to 1.796.

Step 5. Comparing *t* against $t_{cp}$, we find that 1.88 exceeds 1.796, suggesting a statistically significant difference between the background and downgradient population log-means, at least at the 5% level of significance. This means that the downgradient geometric mean concentration — and equivalently for lognormal populations, the median concentration — is statistically greater than the same statistical measure in background. Further, since the downgradient sample log-variance is over three times the magnitude of the background log-variance, it is also probable that the downgradient arithmetic mean is larger than the background arithmetic mean.

Figure 16-3. Benzene Time Series Plot



A note of caution in this example is that the same test run at the $\alpha = 0.01$ level would yield a *non-significant* result, since the upper 99% Student's *t* critical point in that case would be 2.718. The fact that the conclusion differs based on a small change to the significance level ought to prompt review of other *t*-test assumptions. A check of the downgradient sample measurements indicates an upward (non-stationary) trend over the sample collection period (**Figure 16-3**). This reinforces the fact that the *t*-test can be ill-suited for measuring differences between populations when trends over time cause instability in the underlying population means. It might be necessary to either perform a formal test of trend at the downgradient well or to limit the compliance data included in the evaluation only to those most representative of current conditions at the downgradient well (*e.g.*, the last four measurements). ◄

## 16.2 WILCOXON RANK-SUM TEST

### BACKGROUND AND PURPOSE

When the underlying distribution of a data set is unknown and cannot be readily identified as normal or normalized via a transformation, a non-parametric alternative to the two-sample *t*-test is recommended. Probably the best and most practical substitute is the ***Wilcoxon rank-sum test*** (Lehmann, 1975; also known as the two-sample ***Mann-Whitney U test***), which can be used to compare a single compliance well or data group against background. Like many non-parametric methods, the Wilcoxon rank-sum test is based on the ranks of the sample measurements rather than the actual concentrations. Some statistical information contained in the original data is lost when switching to the Wilcoxon test, since it only uses the relative magnitudes of data values.

The benefit is that the ranks can be used to conduct a statistical test even when the underlying population has an unusual form and is non-normal. The parametric *t*-test depends on the population being at least approximately normal; when this is not the case, the critical points of the *t*-test can be highly inaccurate. The Wilcoxon rank-sum test is also a statistically ***efficient*** procedure. That is, when compared to the *t*-test using normally-distributed data especially for larger sample sizes, it performs nearly as well as the *t*-test. Because of this fact, some authors (*e.g.*, Helsel and Hirsch, 2002) have recommended routine use of the Wilcoxon rank-sum even when the parametric *t*-test might be appropriate.

Although a reasonable strategy for larger data sets, one should be careful about automatically preferring the Wilcoxon over the *t*-test on samples as small as those often available in groundwater monitoring. For instance, a Wilcoxon rank-sum test of four samples in each of a background and compliance well and an $\alpha = 0.01$ level of significance can ***never*** identify a significant difference between the two populations. This is true no matter what the sample concentrations are, even if ***all*** four compliance measurements are larger than any of the background measurements. This Wilcoxon test will require at least five samples in at least one of the groups, or a higher level of significance (say $\alpha = 0.05$ or 0.10) is needed.

The Wilcoxon test statistic (*W*) consists of the sum of the ***ranks*** of the compliance well measurements. The rationale of the test is that if the ranks of the compliance data are quite large relative to the background ranks, then the hypothesis that the compliance and background values came from the same population ought to be rejected. Large values of the *W* statistic give evidence of possible

contamination in the compliance well. Small values of *W*, on the other hand, suggest there is little difference between the background and compliance well measurements.

## REQUIREMENTS AND ASSUMPTIONS

The Wilcoxon rank-sum test assumes that both populations being compared follow a common, though unknown, parent distribution under the null hypothesis (Hollander and Wolfe, 1999). Such an assumption is akin to that used in the two-sample pooled variance Student's *t*-test, although the form of the common distribution need not be normal. The Wilcoxon test assumes that both population variances are equal, unlike Welch's *t*-test. Side-by-side box plots of the two data groups can be compared (**Chapter 9**) to examine whether or not the level of variability appears to be approximately equal in both samples. Levene's test (**Chapter 11**) can also be applied as a formal test of heteroscedasticity given its relative robustness to non-normality. If there is a substantial difference in variance between the background and compliance point populations, one remedy is the *Fligner-Policello test* (Hollander and Wolfe, 1999), a more complicated rank-based procedure.

The Wilcoxon procedure as described in the Unified Guidance is generally used as an *interwell* test, meaning that it should be avoided under conditions of significant natural spatial variability. Otherwise, differences between background and compliance point wells identified by the test may be mistakenly attributed to possible contamination, instead of natural differences in geochemistry, *etc.* At small sites, the Wilcoxon procedure can be adapted for use as an *intrawell* test, involving a comparison between intrawell background and more recent measurements from the same well. However, the per-comparison false positive rate in this case should be raised to either $\alpha = 0.05$ or $\alpha = 0.10$. More generally, a significance level of at least 0.05 should be adopted whenever the sample size of either group is no greater than $n = 4$.

In addition to spatial stationarity (*i.e.*, lack of natural spatial variability), the Wilcoxon rank-sum test assumes that the tested populations are stationary *over time*, so that mean levels are not trending upward or downward. As with the *t*-test, if trends are evident in time series plots of the sample data, a formal trend test might need to be employed instead of the Wilcoxon rank-sum, or the scope of the sample may need to be limited to only include data representative of current groundwater conditions.

## HANDLING TIES

When ties are present in a combined data set, adjustments need to be made to the usual Wilcoxon test statistic. Ties will occur in two situations: 1) detected measurements reported with the same numerical value and 2) non-detect measurements with a common RL. Non-detects are considered ties because the actual concentrations are unknown; presumably, every non-detect has a concentration somewhere between zero and the quantitation limit [QL]. Since these measurements cannot be ordered and ranked explicitly, the approximate remedy in the Wilcoxon rank-sum procedure is to treat such values as ties.

One may be able to partially rank the set of non-detects by making use of laboratory-supplied analytical qualifiers. As discussed in **Section 6.3**, there are probable concentration differences between measurements labeled as undetected (*i.e.*, given a "U" qualifier), non-detect (usually reported without a qualifier), or as estimated concentrations (usually labeled with "J" or "E"). One reasonable strategy is to group all U values as the lowest set of ties, other non-detects as a higher set of ties, and to rank all J

and/or E values according to their estimated concentrations. In situations where estimated values for J and E samples are not provided, treat these measurements as the highest group of tied non-detects. Always give the highest ranks to explicitly quantified or estimated concentration measurements. In this way, a more detailed partial ranking of the data will be possible.

Tied observations in the Wilcoxon rank-sum test are handled as follows. All tied observations in a particular group should receive the same rank. This rank called the *midrank* (Lehmann, 1975) is computed as the average of the ranks that would be assigned to a group of ties if the tied values actually differed by a tiny amount and could be ranked uniquely. For example, if the first four ordered observations are all the same, the midrank given to each of these samples would be equal to (1 + 2 + 3 + 4)/4 = 2.5. If the next highest measurement is a unique value, its rank would be 5, and so on until all observations are appropriately ranked. A more detailed example is illustrated in **Figure 16-4**.

Figure 16-4. Computation of Midranks for Groups of Tied Values

| Order | Concentration | Mid-Rank | |
|---|---|---|---|
| 1 | <1 | 1.5 | $\Rightarrow \frac{1}{2}(1+2)$ |
| 2 | <1 | 1.5 | |
| 3 | 1.2 | 3 | |
| 4 | 1.3 | 5 | $\Rightarrow \frac{1}{3}(4+5+6)$ |
| 5 | 1.3 | 5 | |
| 6 | 1.3 | 5 | |
| 7 | 1.5 | 7.5 | $\Rightarrow \frac{1}{2}(7+8)$ |
| 8 | 1.5 | 7.5 | |
| 9 | 1.6 | 9 | |

## HANDLING NON-DETECTS

If either of the samples contains a substantial fraction of non-detect measurements (say more than 20-30%), identification of an appropriate distributional model (*e.g.*, normality) may be difficult, effectively ruling out the use of parametric tests like the *t*-test. Even when a normal or other parametric model can be fit to such left-censored data, a *t*-test cannot be run without imputing estimated values for each non-detect. Past guidance has recommended the Wilcoxon rank-sum test as an alternative to the *t*-test in the presence of non-detects, with all non-detects at a common RL being treated as tied values.

If the combined data set contains a single, common RL, that limit is smaller than any of the detected/quantified values, and the proportion of censored data is small (say no more than 10-15% of the total), it may be reasonable to treat the non-detects as a set of tied values and to apply the Wilcoxon rank-sum test adjusted for ties (described below). More generally, however, the statistical behavior of the Wilcoxon statistic depends on a full and accurate ranking of all the measurements. Groups of left-censored values cannot be ranked with certainty, even if each such measurement possesses a common RL. The problem is compounded in the presence of multiple RLs and/or quantified values less than the

RL(s). What is the relative ranking, for instance, of the pair of measurements (<1, <5)? A higher RL does not guarantee that the second observation is larger in magnitude than the first. A similar uncertainty plagues the pair of values (4, <10). And there is no guarantee either that the pair (<2, <2) is actually tied. One may be able to partially rank the set of non-detects by making use of laboratory-supplied analytical qualifiers as described in the previous section.

Because non-detects generally prevent a complete ranking of the measurements, the Wilcoxon rank-sum test is not recommended for most censored data sets. Instead, a modified version of the Tarone-Ware test (Hollander and Wolfe, 1999) is presented in **Section 16.3**. The Tarone-Ware test is essentially a generalization of the Wilcoxon test specifically designed to accommodate censored values.

## PROCEDURE

Step 1. To conduct the Wilcoxon rank-sum test, first combine the compliance and background data into a single data set. Sort the combined values from smallest to largest, and — if there are no tied values or non-detects with a common RL — rank the ordered values from 1 to $N$. Assume there are $n$ compliance well samples and $m$ background samples so that $N = m + n$. Denote the ranks of the compliance samples by $C_i$ and the ranks of the background samples by $B_i$.

Step 2. If there are groups of tied values (including non-detects with a common RL), form the midranks of the combined data set by assigning to each set of ties the average of the potential ranks the tied members would have been given if they could be uniquely ranked.

Step 3. Sum the ranks of the compliance samples to get the Wilcoxon statistic $W$:

$$W = \Sigma_{i=1}^{n} C_i \qquad [16.5]$$

Step 4. Find the $\alpha$-level critical point of the Wilcoxon test, making use of the fact that the sampling distribution of $W$ under the null hypothesis, $H_0$, can be approximated by a normal curve. By standardizing the statistic $W$ (*i.e.*, subtracting off its mean or expected value and dividing by its standard deviation), the standardized statistic or $z$-score, $Z$, can be approximated by a standard normal distribution. Then an appropriate critical point ($z_{cp}$) can be determined as the upper $(1-\alpha) \times 100$th percentage point of the standard normal distribution, listed in **Table 10-1** in **Appendix D**.

Step 5. To compute $Z$ when there are no ties, first compute the expected value and standard deviation of $W$, given respectively by the following equations:

$$E(W) = \frac{1}{2} n(N+1) \qquad [16.6]$$

$$SD(W) = \sqrt{\frac{1}{12} mn(N+1)} \qquad [16.7]$$

Then compute the approximate $z$-score for the Wilcoxon rank-sum test as:

$$Z = \frac{W - E(W) - 1/2}{SD(W)} \quad [16.8]$$

The factor of 1/2 in the numerator serves as a continuity correction since the discrete distribution of the Wilcoxon statistic $W$ is being approximated by a continuous normal distribution.

Step 6. If there are tied values, compute the expected value of W using [16.6] and the standard deviation of W adjusted for the presence of ties with the equation:

$$SD^*(W) = \sqrt{\frac{mn(N+1)}{12}\left(1 - \Sigma_{i=1}^{g} \frac{t_i^3 - t_i}{N^3 - N}\right)} \quad [16.9]$$

where $g$ equals the number of different groups of tied observations and $t_i$ represents the number of tied values in the $i$th group.

Then compute the approximate z-score for the Wilcoxon rank-sum test as:

$$Z = \frac{W - E(W) - 1/2}{SD^*(W)} \quad (16.10)$$

Step 7. Compare the approximate $z$-score against the critical point, $z_{cp}$. If $Z$ exceeds $z_{cp}$, conclude that the compliance well concentrations are significantly greater than background at the $\alpha$ level of significance. If not, conclude that the null hypothesis of equivalent background and compliance point distributions cannot be rejected.

►EXAMPLE 16-4

The table below contains copper concentrations (ppb) found in groundwater samples at a Western monitoring facility. Wells 1 and 2 denote background wells while Well 3 is a single downgradient well suspected of being contaminated. Calculate the Wilcoxon rank-sum test on these data at the $\alpha$ = .01 level of significance.

| | Copper Concentration (ppb) | | |
|---|---|---|---|
| | Background | | Compliance |
| Month | Well 1 | Well 2 | Well 3 |
| 1 | 4.2 | 5.2 | 9.4 |
| 2 | 5.8 | 6.4 | 10.1 |
| 3 | 11.3 | 11.3 | 14.5 |
| 4 | 7.0 | 11.5 | 16.1 |
| 5 | 7.0 | 10.1 | 21.5 |
| 6 | 8.2 | 9.7 | 17.6 |

SOLUTION

Step 1. Sort the $N$ = 18 observations from least to greatest. Since there are 3 pairs of tied values, compute the midranks as in the table below. Note that $m$ = 12 and $n$ = 6.

Step 2. Compute the Wilcoxon statistic by summing the compliance well ranks, so that $W$ = 84.5.

Step 3. Using $\alpha$ = .01, find the upper 99th percentage point of the standard normal distribution in **Table 10-1** of **Appendix D**. This gives a critical value of $z_{cp}$ = 2.326.

| | Midranks of Copper Concentrations | | |
|---|---|---|---|
| | Background | | Compliance |
| Month | Well 1 | Well 2 | Well 3 |
| 1 | 1 | 2 | 8 |
| 2 | 3 | 4 | 10.5 |
| 3 | 12.5 | 12.5 | 15 |
| 4 | 5.5 | 14 | 16 |
| 5 | 5.5 | 10.5 | 18 |
| 6 | 7 | 9 | 17 |

Step 4. Compute the expected value and adjusted standard deviation of $W$ using equations [16.6] and (16.10), recognizing there are 3 groups of ties with $t_i$ = 2 measurements in each group:

$$E(W) = \frac{1}{2} \cdot 6 \cdot 19 = 57$$

$$SD(W) = \sqrt{\frac{1}{12} \cdot 12 \cdot 6 \cdot (18+1)\left[1 - 3 \cdot \left(\frac{2^3 - 2}{18^3 - 18}\right)\right]} = \sqrt{113.647} = 10.661$$

Then compute the standardized statistic or *z*-score, *Z*, using equation (16.10):

$$Z = \frac{84.5 - 57 - 0.5}{10.661} = 2.533$$

Step 5. Compare the observed *z*-score against the critical point $z_{cp}$. Since $Z = 2.533 > 2.326 = z_{.99}$, there is statistically significant evidence of possible contamination in the compliance well at the $\alpha = .01$ significance level. ◄

## 16.3 TARONE-WARE TWO-SAMPLE TEST FOR CENSORED DATA

### BACKGROUND

In statistical terms, non-detect measurements represent left-censored values, in which the 'true' magnitude is known only to exist somewhere between zero and the RL, i.e., within the concentration interval [0, RL). The uncertainty introduced by non-detects impacts the applicability of other two-sample comparisons like the t-test and Wilcoxon rank-sum test. Because the Student's t-test cannot be run unless a specific magnitude is assigned to each observation, estimated or imputed values need to be assigned to the non-detects. The Wilcoxon procedure requires that every observation be ranked in relation to other values in the combined sample, even though non-detects allow at best only a partial ranking, as discussed in **Section 16.2**.

The Tarone-Ware two-sample test can be utilized to overcome these limitations for many groundwater data with substantial fractions of non-detects along with multiple RLs. Tarone and Ware (1977) actually proposed a family of tests to analyze censored data. One variant of this family is the logrank test, frequently used in survival analysis for right-censored data. Another variant is known as Gehan's generalized Wilcoxon test (Gehan, 1965). The Unified Guidance presents the variant recommended by Tarone and Ware, slightly modified to account for left-censored measurements.

The key benefit of the Tarone-Ware procedure is that it is designed to provide a valid statistical test, even with a large fraction of censored data. As a non-parametric test, it does not require normally-distributed observations. In addition, non-detects do not have to be imputed or even fully ranked. Instead, for each detected concentration (*c*), a simple count needs to be made within each sample of the number of detects and non-detects no greater in magnitude than *c*. These counts are then combined to form the Tarone-Ware statistic.

### REQUIREMENTS AND ASSUMPTIONS

The null hypothesis ($H_0$) under the Tarone-Ware procedure assumes that the populations in background and the compliance well being tested are identical. This implies that the variances in the two distributions are the same, thus necessitating a check of equal variances. With many non-detect data sets, it can be very difficult to formally test for heteroscedasticity. Often the best remedy is to make an informal, visual check of variability using side-by-side box plots (**Chapter 9**), setting each non-detect to half its RL.

The Tarone-Ware test will typically be used as an *interwell* test, meaning that it should be avoided under conditions of significant natural spatial variability. In addition, the tested populations should be stationary *over time*, so that mean levels are not trending upward or downward. Both assumptions can be more difficult to verify with censored data. Spatial variation can sometimes be checked with a non-parametric Kruskal-Wallis analysis of variance (**Chapter 17**). Trends with censored data can be identified with the Mann-Kendall test (**Chapter 14**).

As with other two-sample tests, if a trend is identified in one or both samples, a formal trend test may be needed instead of the Tarone-Ware, or the scope of the sample may need to be limited to only include data representative of current groundwater conditions.

Because the Tarone-Ware test presented in the Unified Guidance depends on counts of observations with magnitudes no greater than each detected concentration, and in that sense generalizes the ranking process used by the Wilcoxon rank-sum procedure, it is recommended that estimated concentrations (*i.e.*, sample measurements assigned unique magnitudes but labeled with qualifiers "J" or "E") be treated as detections for the purpose of computing the Tarone-Ware statistic. Such observations provide valuable statistical information about the relative ranking of each censored sample, even if estimated concentrations possess larger measurement uncertainty than fully quantified values.

## PROCEDURE

Step 1. To compare a background data set against a compliance well using the Tarone-Ware test, first combine the two samples. Locate and sort the $k$ distinct detected values and label these as:

$$w_{(1)} \langle w_{(2)} \langle \ldots \langle w_{(k-1)} \langle w_{(k)}$$

Note that the set of $w$'s will not include any RLs from non-detects. Also, if two or more detects are tied, $k$ will be less than the total number of detected measurements.

Step 2. For the combined sample, count the number of observations (described by Tarone & Ware as 'at risk') for each distinct detected concentration. That is, for $i = 1,\ldots,k$, let $n_i$ = the number of detected values no greater than $w_{(i)}$ plus the number of non-detects with RLs no greater than $w_{(i)}$. Also let $d_i$ = the number of detects with concentration equal to $w_{(i)}$. This value will equal 1 unless there are multiple detected values with the same reported concentration.

Step 3. For the compliance sample, count the observations 'at risk', much as in Step 2. For $i = 1$ to $k$, let $n_{i2}$ = the number of detected compliance values no greater than $w_{(i)}$ plus the number of compliance point non-detects with RLs no greater than $w_{(i)}$. Also let $d_{i2}$ = the number of compliance point detects with concentration equal to $w_{(i)}$. Note that $d_{i2} = 0$ if $w_{(i)}$ represents a detected value from background. Also compute $n_{i1}$, the number 'at risk' in the background sample.

Step 4. For $i = 1$ to $k$, compute the expected number of compliance point detections using the formula:

$$E_{i2} = d_i n_{i2} / n_i \qquad (16.11)$$

Also compute the variance of the number of compliance point detections, using the equation:

$$V_{i2} = \frac{d_i\left(n_i - d_i\right)n_{i1}n_{i2}}{n_i^2\left(n_i - 1\right)} \tag{16.12}$$

Note in equation (16.12) that if $n_i = 1$ for the smallest detected value, the numerator of $V_{i2}$ will necessarily equal zero (since $d_i = 1$ in that case), so compute $V_{i2} = 0$.

Step 5. Construct the Tarone-Ware statistic (*TW*) with the equation:

$$TW = \frac{\sum_{i=1}^{k}\sqrt{n_i}\left(d_{i2} - E_{i2}\right)}{\sqrt{\sum_{i=1}^{k} n_i V_{i2}}} \tag{16.13}$$

Step 6. Find the $\alpha$-level critical point of the Tarone-Ware test, making use of the fact that the sampling distribution of *TW* under the null hypothesis, $H_0$, is designed to approximately follow a standard normal distribution. An appropriate critical point ($z_{cp}$) can be determined as the upper $(1-\alpha) \times 100$th percentage point of the standard normal distribution, listed in **Table 10-1** of **Appendix D**.

Step 7. Compare *TW* against the critical point, $z_{cp}$. If *TW* exceeds $z_{cp}$, conclude that the compliance well concentrations are significantly greater than background at the $\alpha$ level of significance. If not, conclude that the null hypothesis of equivalent background and compliance point distributions cannot be rejected.

### ►EXAMPLE 16-5

A heavily industrial site has been historically contaminated with tetrachloroethylene [PCE]. Using the Tarone-Ware procedure at an $\alpha = .05$ significance level, test the following PCE measurements collected from one background and one compliance well.

| PCE (ppb) | |
|---|---|
| Background | Compliance |
| <4 | 6.4 |
| 1.5 | 10.9 |
| <2 | 7 |
| 8.7 | 14.3 |
| 5.1 | 1.9 |
| <5 | 10.0 |
| | 6.8 |
| | <5 |

#### SOLUTION

Step 1. Combine the background and compliance point samples. List and sort the distinct detected values (as in the table below), giving $k = 10$. Note that the 4 non-detects comprise 28% of the combined data.

Step 2. Compute the number of measurements ($n_i$) in the combined sample 'at risk' for each distinct detected value ($w_{(i)}$), indexed from $i = 1,\ldots, 10$, by adding the number of detects and non-

detects no greater than $w_{(i)}$, as listed in column 6 of the table below. Also list in column 3 the number of detected values ($d_i$) exactly equal to $w_{(i)}$.

Step 3. For the compliance point sample, compute the number ($n_{i2}$) 'at risk' for each distinct detected value, as listed in column 5 below. Also compute the number ($n_{i1}$) 'at risk' for the background sample (column 4) and the number of compliance point measurements exactly equal to $w_{(i)}$ (column 2).

Step 4. Use equations (16.11) and (16.12) to compute the expected value ($E_{i2}$) and variance ($V_{i2}$) of the number of compliance point detections at each $w_{(i)}$ (columns 7 and 8 below).

| $w_{(i)}$ | $d_{i2}$ | $d_i$ | $n_{i1}$ | $n_{i2}$ | $n_i$ | $E_{i2}$ | $V_{i2}$ |
|---|---|---|---|---|---|---|---|
| 1.5 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| 1.9 | 1 | 1 | 1 | 1 | 2 | 0.5 | 0.25 |
| 5.1 | 0 | 1 | 5 | 2 | 7 | 0.2857 | 0.2041 |
| 6.4 | 1 | 1 | 5 | 3 | 8 | 0.375 | 0.2344 |
| 6.8 | 1 | 1 | 5 | 4 | 9 | 0.4444 | 0.2469 |
| 7.0 | 1 | 1 | 5 | 5 | 10 | 0.5 | 0.25 |
| 8.7 | 0 | 1 | 6 | 5 | 11 | 0.4545 | 0.2479 |
| 10.0 | 1 | 1 | 6 | 6 | 12 | 0.5 | 0.25 |
| 10.9 | 1 | 1 | 6 | 7 | 13 | 0.5385 | 0.2485 |
| 14.3 | 1 | 1 | 6 | 8 | 14 | 0.5714 | 0.2449 |

Step 5. Calculate the Tarone-Ware statistic (TW) using equation (16.13):

$$TW = \frac{\sqrt{1}\cdot(0-0)+\sqrt{2}\cdot(1-0.5)+\sqrt{7}\cdot(0-.2857)+\ldots+\sqrt{14}\cdot(1-.5714)}{\sqrt{1\cdot 0+2\cdot .25+7\cdot .2041+\ldots+14\cdot .2449}} = 1.85$$

Step 6. Determine the 0.05 level critical point from **Table 10-1** in **Appendix D** as the upper 95th percentage point from a standard normal distribution. This gives $z_{cp} = 1.645$.

Step 7. Compare the Tarone-Ware statistic against the critical point. Since $TW = 1.85 > 1.645 = z_{cp}$, conclude that the PCE concentrations are significantly greater at the compliance well than in background at the 5% significance level. ◄

*This page intentionally left blank*

# CHAPTER 17. ANOVA, TOLERANCE LIMITS, AND TREND TESTS



This chapter describes two statistical procedures — analysis of variance [ANOVA] and tolerance limits — explicitly allowed within §264.97(h) and §258.53(g) for use in groundwater monitoring. The Unified Guidance does not generally recommend either technique for formally making regulatory decisions about compliance wells or regulated units, instead focusing on prediction limits, control charts, and confidence intervals. But both ANOVA and tolerance tests are standard statistical procedures that can be adapted for a variety of uses. ANOVA is particularly helpful in both identifying on-site spatial variation and in sometimes aiding the computation of more effective and statistically powerful intrawell prediction limits (see **Chapters 6** and **13** for further discussion).

This chapter also presents selected trend tests as an alternative statistical method that can be quite useful in groundwater detection monitoring, particularly when groundwater populations are not *stationary* over time. Although trend tests are not explicitly listed within the RCRA regulations, they possess advantages in certain situations and can meet the performance requirements of §264.97(i) and §258.53(h). They can also be helpful during diagnostic evaluation and establishment of historical background (**Chapter 5**) and in verifying key statistical assumptions (**Chapter 14**).

## 17.1 ANALYSIS OF VARIANCE [ANOVA]

### 17.1.1 ONE-WAY PARAMETRIC F-TEST

#### BACKGROUND AND PURPOSE

The parametric one-way ANOVA is a statistical procedure to determine whether there are statistically significant differences in mean concentrations among a set of wells. In groundwater applications, the question of interest is whether there is potential contamination at one or more compliance wells when compared to background. By finding a significant difference in means and specifically ***higher*** average concentrations at one or more compliance wells, ANOVA results can sometimes be used to identify unacceptably high contaminant levels in the absence of natural spatial variability.

Like the two-sample *t*-test, the one-way ANOVA is a comparison of population means. However, the one-way parametric ANOVA is a comparison of *several* populations, not just two: one set of

background data versus at least two compliance wells. The *F*-statistic that forms the heart of the ANOVA procedure is actually an extension of the *t*-statistic; an *F*-statistic formed in a comparison of only two datasets reduces to the square of the usual pooled variance Student's *t*-statistic. Like the *t*-statistic, the *F*-statistic is a ratio of two quantities. The numerator is a measure of the average squared difference observed between the pairs of sample means, while the denominator represents the average variability found in each well group.

Under the null hypothesis that all the wells or groups have the same population mean, the *F*-statistic follows the *F*-distribution. Unlike the *t*-distribution with a single degrees of freedom *df*, there are two *df* quantities associated with *F*. One is for the numerator and the other for the denominator. When critical points are needed from the *F*-distribution, one must specify both degrees of freedom values.

Computation of the *F*-statistic is only the first step of the full ANOVA procedure, when used as a formal compliance test. It can only determine whether *any* significant mean difference exists between the possible pairs of wells or data groups, and not whether or what specific compliance wells differ from background. To accomplish this latter task when a significant *F*-test is registered, individual tests between each compliance well and background needs to be conducted, known as individual *post-hoc comparisons* or *contrasts*. These individual tests are a specially constructed series of *t*-tests, with critical points chosen to limit the *test-wise* or *experiment-wise* false positive rate.

## REQUIREMENTS AND ASSUMPTIONS

The parametric ANOVA assumes that the data groups are normally-distributed with constant variance. This means that the group *residuals* should be tested for normality (**Chapter 10**) and that the groups have to be tested for equality of variance, perhaps with Levene's test (**Chapter 11**). Since the *F*-test used in the one-way ANOVA is reasonably robust to small departures from normality, the first of these assumptions turns out to be less critical than the second. Research (Milliken and Johnson, 1984) has shown that the statistical power of the *F*-test is strongly affected by inequality in the population variances. A noticeable drop in power is seen whenever the ratio of the largest to smallest group variance is at least 4. A severe drop in power is found whenever the ratio of the largest to smallest group variance is at least a factor of 10. These ratios imply that the *F*-test will lose some statistical power if any of the group population standard deviations is at least twice the size of any other group's standard deviation, and that the power will be greatly curtailed if any standard deviation is at least 3 times as large as any other group's.

If the hypothesis of equal variances is rejected or if the group residuals are found to violate an assumption of normality (especially at the .01 significance level or less), one should consider a transformation of the data, followed by testing of the ANOVA assumptions on the transformed scale. If the residuals from the transformed data still do not satisfy normality or if there are too many non-detect measurements to adequately test the assumptions, a non-parametric ANOVA (called the *Kruskal-Wallis test*) using the ranks of the observations is recommended instead (see **Section 17.1.2**).

Since ANOVA is inherently an interwell statistical method, a critical point in using ANOVA for compliance testing is that the well field should exhibit *minimal spatial variability*. Interwell tests also require the groundwater well populations to be spatially *stationary*, so that absent a release the population well means are stable over time. Because spatial variation is frequently observed in many

groundwater constituents, especially for common inorganic constituents and some metals, ANOVA may not be usable as compliance testing tool. Yet it can be utilized on the same data sets to help *identify* the presence of spatial variability. In this capacity, the same procedure and formulas are utilized as described below (with the exception of the post-hoc contrasts, which are unnecessary for assessing spatial variation). The results are then employed to guide the appropriate choice of a compliance test (*e.g.*, intrawell or interwell prediction limits).

For formal ANOVA testing under §264.97(i) and §258.53(h), the *experiment-wise* or *test-wise* false positive rate ($\alpha$) needs to be at least 5% during any statistical evaluation for each constituent tested. Furthermore, the individual *post-hoc contrasts* used to test single compliance wells against background need to be run at a significance level of at least $\alpha^* = 1\%$ per well. Combined, these regulatory constraints imply that if there are more than five post-hoc contrasts that need to be tested (*i.e.*, more than 5 compliance wells are included in the ANOVA test), the overall, maximal false positive rate of the procedure will tend to be greater, and perhaps substantially so, than 5%. Also, since $\alpha = 5\%$ is the minimum significance level per monitoring constituent, running multiple ANOVA procedures to accommodate a list of constituents will lead to a minimum *site-wide false positive rate* [SWFPR] greater than the Unified Guidance recommended target of 10% per statistical evaluation.

In addition, if a contaminated compliance well exists but too many uncontaminated wells are also included in the same ANOVA, the *F*-statistic may result in a non-significant outcome. Performing ANOVA with more than 10 to 15 well groups can "swamp" the procedure, causing it to lose substantial power. It therefore will be necessary to consider one of the retesting strategies described in **Chapters 18** and **20** as an alternative to ANOVA in the event that either the expected false positive rate is too large, or if more than a small number of wells need to be tested.

Another drawback to the one-way ANOVA is that the *F*-test accounts for *all possible paired comparisons* among the well groups. In some cases, the *F*-statistic may be *significant* even though all of the contrasts between compliance wells and background are *non-significant*. This does not mean that the *F*-test has necessarily registered a false positive. Rather, it may be that two of the compliance wells significantly differ from each other, but neither differs from background. This could happen, for instance, if a compliance well has a lower mean concentration than background while other compliance wells have near background means. The *F*-test looks for all possible differences between pairs of well groups, not just those comparisons against background.

In order to run a valid one-way *F*-test, a minimum number of observations are needed. Denoting the number of data groups by $p$, at least $p > 2$ groups must be compared (*e.g.*, two or more compliance wells versus background). Each group should have at least three to four statistically independent observations and the total sample size, $N$, should be large enough so that $N–p > 5$. As long as $p \geq 3$ and there are at least 3 observations per well, this last requirement will always be met. But the statistical power of an ANOVA to identify differences in population means tends to be minimal unless there are at least 4 or more observations per data group. It is also helpful to have at least 8 measurements in background for the test.

Similarly to the two-sample *t*-test, it may be very difficult to verify that the measurements are statistically independent with only a handful of observations per well. One should additionally ensure that the samples are collected far enough apart in time to avoid significant *autocorrelation* (see **Chapter 14** for further discussion). A periodic check of statistical independence in each may be possible after a

few testing periods, when enough data has been collected to enable a statistical assessment of this assumption.

## PROCEDURE

Step 1. Combine all the relevant background data collected from multiple wells into one group. These wells should have insignificant mean differences under prior ANOVA testing. If the regulated unit has ($p$–1) compliance wells, there will then be a total of $p$ data groups. Because there may be different numbers of observations per well, denote the sample size of the $i$th group by $n_i$ and the total number of data points across all groups by $N$.

Step 2. Denote the observations in the $i$th well group by $x_{ij}$ for $i$ = 1 to $p$ and $j$ = 1 to $n_i$. The first subscript designates the well, while the second denotes the $j$th value in the $i$th well. Then compute the mean of each well group along with the overall (grand) mean of the combined dataset using the following formulas:

$$\bar{x}_{i\bullet} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \qquad [17.1]$$

$$\bar{x}_{\bullet\bullet} = \frac{1}{N}\sum_{i=1}^{p}\sum_{j=1}^{n_i} x_{ij} \qquad [17.2]$$

Step 3. Compute the sum of squares of differences between the well group means and the grand mean, denoted $SS_{wells}$:

$$SS_{wells} = \sum_{i=1}^{p} n_i\left(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet}\right)^2 = \sum_{i=1}^{p} n_i\bar{x}_{i\bullet}^2 - N\bar{x}_{\bullet\bullet}^2 \qquad [17.3]$$

The formula on the far right is usually the most convenient for calculation. This sum of squares has ($p$–1) degrees of freedom associated with it and is a measure of the variability ***between*** wells. It constitutes the numerator of the $F$-statistic.

Step 4. Compute the corrected total sum of squares, denoted by $SS_{total}$:

$$SS_{total} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}_{\bullet\bullet}\right)^2 = \sum_{i=1}^{p}\sum_{j=1}^{n_i} x_{ij}^2 - N\bar{x}_{\bullet\bullet}^2 \qquad [17.4]$$

The far right equation is convenient for calculation. This sum of squares has ($N$–1) degrees of freedom associated with it and is a measure of the variability in the entire dataset. In fact, if $SS_{total}$ is divided by ($N$–1), one gets the overall sample variance.

Step 5. Compute the sum of squares of differences between the observations and the well group means. This is known as the within-wells component of the total sum of squares or, equivalently, as the sum of squares due to error. It is easiest to obtain by subtraction using the far right side of equation [17.5] and is denoted $SS_{error}$:

$$SS_{error} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}_{i\bullet}\right)^2 = SS_{total} - SS_{wells} \quad [17.5]$$

$SS_{error}$ is associated with ($N$–$p$) degrees of freedom and is a measure of the variability ***within*** well groups. This quantity goes into the denominator of the $F$-statistic.

Step 6. Compute the mean sum of squares for both the between-wells and within-wells components of the total sum of squares, denoted by $MS_{wells}$ and $MS_{error}$. These quantities are simply obtained by dividing each sum of squares by its corresponding degrees of freedom:

$$MS_{wells} = SS_{wells} \big/ \left(p - 1\right) \quad [17.6]$$

$$MS_{error} = SS_{error} \big/ \left(N - p\right) \quad [17.7]$$

Step 7. Compute the $F$-statistic by forming the ratio between the mean sum of squares for wells and the mean sum of squares due to error, as in **Figure 17-1**. This layout is known as a one-way parametric ANOVA table and illustrates the sum of squares contribution to the total variability, along with the corresponding degrees of freedom, the mean squares components, and the final $F$-statistic calculated as $F = MS_{wells}/MS_{error}$. Note that the first two rows of the one-way table sum to the last row.

Figure 17-1. One-Way Parametric ANOVA Table

| Source of Variation | Sums of Squares | Degrees of Freedom | Mean Squares | $F$-Statistic |
|---|---|---|---|---|
| Between Wells | $SS_{wells}$ | $p$–1 | $MS_{wells} = SS_{wells}/(p–1)$ | $F = MS_{wells}/MS_{error}$ |
| Error (within wells) | $SS_{error}$ | $N$–$p$ | $MS_{error} = SS_{error}/(N–p)$ | |
| Total | $SS_{total}$ | $N$–1 | | |

Step 8. To test the hypothesis of equal means for all $p$ wells, compare the $F$-statistic in **Figure 17-1** to the $\alpha$-level critical point found from the $F$-distribution with ($p$–1) and ($N$–$p$) degrees of freedom in **Table 17-1** of **Appendix D**. $\alpha$ is usually set at 5%, so that the needed comparison value equals the upper 95th percentage point of the $F$-distribution. The numerator (*p-1*) and denominator (*N-p*) degrees of freedom for the $F$-statistic are obtained from the above table. If the observed $F$-statistic exceeds the critical point ($F_{.95,\ p–1,\ N–p}$), reject the hypothesis of equal well group population means. Otherwise, conclude that there is insufficient evidence of a significant difference between the concentrations at the $p$ well groups and thus no evidence of potential contamination in any of the compliance wells.

Step 9. In the case of a significant $F$-statistic that exceeds the critical point in Step 8, determine which compliance wells have elevated concentrations compared to background. This is done by comparing each compliance well individually against the background measurements. Tests to assess concentration differences between a pair of well groups are called ***contrasts*** in a multiple comparisons ANOVA framework. Since the contrasts are a series of individual

statistical tests, each run at a fixed significance level $\alpha^*$, the Type I error accumulates across the tests as the number of contrasts increases.

To keep the overall false positive rate close to the targeted rate of 5%, the individual contrasts should be set up as follows: Given ($p$–1) separate background-compliance contrasts, if ($p$–1) ≤ 5, run each contrast at a significance level equal to $\alpha^* = .05/(p–1)$. However, if ($p$–1) > 5, run each contrast at a significance level equal to $\alpha^* = .01$. Note that when there are more than 5 compliance wells, this last provision will tend to raise the overall false positive rate above 5%.

Step 10. Denote the background data set as the first well group, so that the number of background samples is equal to $n_b$. Then for each of the remaining ($p$–1) well groups, compute the standard error of the difference between each compliance well and background:

$$SE_i = \sqrt{MS_{error} \cdot \left( \frac{1}{n_b} + \frac{1}{n_i} \right)} \qquad [17.8]$$

Note that $MS_{error}$ is taken from the one-way ANOVA table in **Figure 17-1**. The standard error here is an extension of the standard error of the difference involving the pooled variance in the Student's $t$-test of **Chapter 16**.

Step 11. Treat the background data as the first well group with the average background concentration equal to $\bar{x}_b$. Compute the Bonferroni $t$-statistic for each of the ($p$–1) compliance wells from $i$ = 2 to $p$, dividing the standard error in Step 10 into the difference between the average concentration at the compliance well and the background average, as shown below:

$$t_i = \left( \bar{x}_i - \bar{x}_b \right) / SE_i \qquad [17.9]$$

Step 12. The Bonferroni $t$-statistic in equation [17.9] is a type of $t$-test. Since the estimate of variability used in equation [17.8] has ($N$–$p$) degrees of freedom, the critical point can be determined from the Student's $t$-distribution in **Table 16-1** of **Appendix D**. Let the Bonferroni critical point ($t_{cp}$) be equal to the upper $(1–\alpha^*) \times 100$th percentage point of the $t$-distribution with ($N$–$p$) degrees of freedom.

Step 13. If any of the Bonferroni $t$-statistics ($t_i$) exceed the critical point $t_{cp}$, conclude that these compliance wells have population mean concentrations significantly greater than background and thus exhibit evidence of possible contamination. Compliance wells for which the Bonferroni $t$-statistic does not exceed $t_{cp}$ should be regarded as similar to background in mean concentration level.

### ►EXAMPLE 17-1

Lead concentrations in ground water at two background and four compliance wells were tested for normality and homoscedasticity. These data were found to be best fit by a lognormal distribution with approximately equal variances. The two background wells also indicated insignificant log mean differences. The natural logarithms of these lead values are shown in the table below. Use the one-way parametric ANOVA to determine whether there are any significant concentration increases over background in any of the compliance wells.

| Log(Lead) log(ppb) | | | | | | |
|---|---|---|---|---|---|---|
| | Background | | Compliance | | | |
| Date | Well 1 | Well 2 | Well 3 | Well 4 | Well 5 | Well 6 |
| Jan 1995 | 4.06 | 3.83 | 4.61 | 3.53 | 4.11 | 4.42 |
| Apr 1995 | 3.99 | 4.34 | 5.14 | 4.54 | 4.29 | 5.21 |
| Jul 1995 | 3.40 | 3.47 | 3.67 | 4.26 | 5.50 | 5.29 |
| Oct 1995 | 3.83 | 4.22 | 3.97 | 4.42 | 5.31 | 5.08 |
| Well Mean | 3.82 | 3.96 | 4.35 | 4.19 | 4.80 | 5.00 |
| Well SD | 0.296 | 0.395 | 0.658 | 0.453 | 0.704 | 0.143 |
| | $\bar{x}_{BG}$ = 3.89 | $s_{BG}$ = 0.333 | Grand Mean = 4.35 | | | |

SOLUTION

Step 1. Combine the two background wells into one group, so that the background average becomes 3.89 log(ppb). Then $n_b = 8$, while $n_i = 4$ for each of the other four well groups. Note that the total sample size is $N = 24$ and $p = 5$.

Step 2. Compute the (overall) grand mean and the sample mean concentrations in each of the well groups using equations [17.1] and [17.2]. These values are listed (along with each group's standard deviation) in the above table.

Step 3. Compute the sum of squares due to well-to-well differences using equation [17.3]:

$$SS_{wells} = \left[8 \cdot (3.89)^2 + 4 \cdot (4.35)^2 + \ldots + 4 \cdot (5.00)^2\right] - 24 \cdot (4.35)^2 = 4.289$$

This quantity has (5–1) = 4 degrees of freedom.

Step 4. Compute the corrected total sum of squares using equation [17.4] with ($N$–1) = 23 *df*:

$$SS_{total} = \left[(4.06)^2 + \ldots + (5.08)^2\right] - 24 \cdot (4.35)^2 = 8.934$$

Step 5. Obtain the within-well or error sum of squares by subtraction using equation [17.5]:

$$SS_{error} = 8.934 - 4.289 = 4.646$$

This quantity has ($N$–$p$) = 24–5 = 19 degrees of freedom.

Step 6. Compute the mean sums of squares using equations [17.6] and [17.7]:

$$MS_{wells} = 4.289/4 = 1.072$$

$$MS_{error} = 4.646/19 = 0.245$$

Step 7. Construct the F-statistic and the one-way ANOVA table, using **Figure 17-1** in **Appendix D** as a guide:

| Source of Variation | Sums of Squares | Degrees of Freedom | Mean Squares | *F*-Statistic |
|---|---|---|---|---|
| Between Wells | 4.289 | 4 | 1.072 | $F$ = 1.072/0.245 |
| Error (within wells) | 4.646 | 19 | 0.245 | = 4.39 |
| Total | 8.934 | 23 | | |

Step 8. Compare the observed *F*-statistic of 4.39 against the critical point taken as the upper 95th percentage point from the *F*-distribution with 4 and 19 degrees of freedom. Using **Table 17-1**, this gives a value of $F_{.95,4,19}$ = 2.90. Since the *F*-statistic exceeds the critical point, the hypothesis of equal well means is rejected, and post-hoc Bonferroni *t*-test comparisons should be conducted.

Step 9. Determine the number of individual contrasts needed. With four compliance wells, ($p$–1) = 4 comparisons need to be made against background. Therefore, run each Bonferroni *t*-test at the $\alpha^*$ = .05/4 = .0125 level of significance.

Step 10. Compute the standard error of the difference between each compliance well average and the background mean using equation [17.8]. Since the number of observations is the same in each compliance well, the standard error in all four cases will be equal to:

$$SE_i = \sqrt{0.245\left(\frac{1}{8}+\frac{1}{4}\right)} = 0.303$$

Step 11. Compute the Bonferroni *t*-statistic for each compliance well using equation [17.9]:

$$\text{Well 3:}\quad t_2 = \left(4.35 - 3.89\right) \Big/ 0.303 = 1.52$$
$$\text{Well 4:}\quad t_3 = \left(4.19 - 3.89\right) \Big/ 0.303 = 0.99$$
$$\text{Well 5:}\quad t_4 = \left(4.80 - 3.89\right) \Big/ 0.303 = 3.00$$
$$\text{Well 6:}\quad t_5 = \left(5.00 - 3.89\right) \Big/ 0.303 = 3.66$$

Note that because Wells 1 and 2 jointly constitute background, the subscripts above correspond to the well groups and not the actual well numbers. Thus, subscript 2 in the Bonferroni *t*-statistic corresponds to Well 3, subscript 3 corresponds to Well 4, and so forth.

Step 12. Look up the critical point from the *t*-distribution in **Table 16-1** of **Appendix D** using a significance level of $\alpha^*$ = .0125 and ($N$–$p$) = 19 *df*. This gives $t_{cp}$ = 2.433.

Step 13. Compare each Bonferroni *t*-statistic from Step 11 against the critical point from Step 12. Because the *t*-statistics at compliance wells 5 and 6 both exceed 2.433, while those at wells 3 and 4 do not, conclude that the population averages in compliance wells 5 and 6 are significantly higher than background. ◄

## 17.1.2 KRUSKAL-WALLIS TEST

### BACKGROUND AND PURPOSE

The parametric one-way ANOVA makes a key assumption that the data residuals are normally-distributed. If this assumption is inappropriate or cannot be tested because of a large fraction of non-detects, a non-parametric ANOVA can be conducted using the ***ranks*** of the observations rather than the original observations. In **Chapter 16**, the Wilcoxon rank-sum test is presented as a non-parametric alternative to the Student's *t*-test when comparing two groups. The Kruskal-Wallis test is offered as a non-parametric alternative to the one-way *F*-test when several groups need to be simultaneously compared, for instance when assessing patterns of spatial variability. Instead of a test of means, the Kruskal-Wallis tests differences among average population ranks equivalent to the ***medians***.

The Kruskal-Wallis test statistic, *H*, does not have the intuitive form of the Student's *t*-test. Under the null hypothesis that all the sample measurements come from identical parent populations, the Kruskal-Wallis statistic follows the well-known ***chi-square*** statistical distribution. Critical points for the Kruskal-Wallis test can be found as upper percentage points of the chi-square ( $\chi^2_{1-\alpha,df}$ ) distribution in **Table 17-2** of **Appendix D**.

If *H* indicates a significant difference between the populations, individual post-hoc comparisons between each compliance well and background need to be conducted if the Kruskal-Wallis is being used for formal compliance testing. Post-hoc contrasts are not generally necessary for identifying spatial variability. Rather than Bonferroni *t*-statistics, contrasts are based on the data ranks and approximately follow a standard normal distribution. The critical points for these contrasts can be obtained from the standard normal distribution in **Table 10-1** of **Appendix D**.

### REQUIREMENTS AND ASSUMPTIONS

While the Kruskal-Wallis test does not require the underlying populations to be normally-distributed, statistical independence of the data is still assumed. Under the null hypothesis of no difference among the groups, the observations are assumed to arise from identical distributions with equal population variances (Hollander and Wolfe, 1999). However, the form of the distribution need not be specified.

A non-parametric ANOVA can be used in any situation that the parametric ANOVA can be used. The minimum data requirements are similar: the sample size for each group in the Kruskal-Wallis procedure should generally be at least four to five observations per group. Despite this similarity, it is often true that non-parametric tests require larger sample sizes than their parametric test counterparts to ensure a similar level of statistical power or ***efficiency***. Non-parametric tests make fewer assumptions concerning the underlying data distribution and so more observations are often needed to make the same judgment that would be rendered by a parametric test. However, the greater efficiency of parametric tests is only achieved when the parent population follows certain known statistical distributions. When the distribution is unknown, non-parametric tests may have much greater power than their parametric counterparts.

Even when a known statistical distribution is considered, rank-based non-parametric tests like the Wilcoxon rank-sum and Kruskal-Wallis often perform reasonably well compared to the t-test and ANOVA. The relative *efficiency* of two procedures is defined as the ratio of the sample sizes needed by each to achieve a certain level of power against a specified alternative hypothesis. As sample sizes get larger, the efficiency of the Kruskal-Wallis test relative to the parametric ANOVA approaches a limit that depends on the underlying distribution of the data, but is always at least 86 percent. This means roughly that, in the worst case, if 86 measurements are available for a parametric ANOVA, only 100 sample values are needed to have an equivalently powerful Kruskal-Wallis test. In many cases, the increase in sample size necessary to match the power of a parametric ANOVA is much smaller or not needed at all. The efficiency of the Kruskal-Wallis test is 95% if the underlying data are really normal, and can be much larger than 100% in other cases (*e.g.*, it is 150% if the data residuals follow a distribution called the *double exponential*). When the efficiency exceeds 100%, the Kruskal-Wallis actually needs fewer observations than the parametric ANOVA to achieve a certain power.

These results imply that the Kruskal-Wallis test is reasonably powerful for detecting concentration differences despite the fact that the original data have been replaced by their ranks. The test can be used with fair success even when the data are normally-distributed and the Kruskal-Wallis is not needed. When the data are not normal or a normalizing transformation cannot be found, the Kruskal-Wallis procedure tends to be more powerful for detecting differences than the usual parametric approach.

## ADJUSTING FOR TIED OBSERVATIONS

The Kruskal-Wallis procedure will frequently be used when the sample data contain a significant fraction of non-detects. However, the presence of non-detects prevents a unique and complete ranking of the concentration values since the exact values of non-detects are unknown.

To address this problem, two steps are necessary. Since they cannot be uniquely ranked, all non-detects are to be treated statistically as 'tied' values. This is an imperfect remedy, since non-detects represent left-censored values and are not necessarily tied. Unfortunately, there is no straightforward, easily implemented alternative to the Kruskal-Wallis for comparing three or more groups containing left-censored observations, unlike the Tarone-Ware alternative to the Wilcoxon rank-sum test discussed in **Chapter 16**. So in the presence of ties (*e.g.*, non-detects or quantified concentrations rounded to the same value), all tied observations should receive the same midrank (discussed in **Section 16.3**). This rank is computed as the *average* of the ranks that would be given to each group of ties if the tied values actually differed by a tiny amount and could be ranked.

To account for multiple reporting limits, all non-detects should be treated as if censored at the highest reporting limit [RL] in the overall sample. Thus, a non-detect reported as <5 would be treated as 'tied' with a non-detect reported as <1, due to the impossibility of knowing which value is actually larger. The only exception to this strategy is when laboratory qualifiers can be used to rank some non-detects as probably greater in magnitude than others. A reasonable strategy discussed in **Section 16.3** is to group all "U" values as the lowest set of ties, other non-detects as a higher set of ties, and to rank all "J" and/or "E" values according to their estimated concentrations. In situations where estimated values for J and E samples are not provided, treat these measurements as the highest group of tied non-detects. Always give the highest ranks to explicitly quantified or estimated concentration measurements.

The second step for handling ties is to compute the Kruskal-Wallis statistic as described below, using for each tied value its corresponding midrank. Then an adjustment to the Kruskal-Wallis statistic needs to be made to account for the presence of ties. This adjustment requires computation of the formula:

$$H^* = H \Bigg/ \left[ 1 - \left( \sum_{i=1}^{g} \frac{t_i^3 - t_i}{N^3 - N} \right) \right] \qquad [17.10]$$

where $g$ equals the number of distinct groups of tied observations, $N$ is the total sample size across all groups, and $t_i$ is the number of observations in the $i$th tied group. Unless there are a substantial number of ties in the overall dataset, the adjustment in equation [17.10] will tend to be small. Still, it is important to properly account for the presence of tied values.

## PROCEDURE

Step 1. To run the Kruskal-Wallis test, denote the total sample size across all well groups by $N$. Temporarily combine all the data into one group and rank the observations from smallest to largest. Treat all non-detects as tied at the lowest possible concentration value, unless using lab qualifiers to distinguish between 'undetected' and other non-detects. Combine all background wells into a single group where appropriate. Denote this set of background data as group 1. Then let $R_{ij}$ denote the $j$th rank from the $i$th well group, and let $k$ equal the total number of groups (*i.e.*, one group of background values and ($k$–1) groups of compliance wells).

Step 2. Compute the sum of the ranks and the average rank in each well group, letting $n_i$ equal the sample size in the $i$th group and using the following formulas:

$$R_{i\bullet} = \sum_{j=1}^{n_i} R_{ij} \qquad [17.11]$$

$$\bar{R}_{i\bullet} = \frac{1}{n_i} R_{i\bullet} \qquad [17.12]$$

Step 3. Calculate the Kruskal-Wallis test statistic $H$ and the adjustment for ties, if necessary, using equation [17.10], where H is given by:

$$H = \left[ \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_{i\bullet}}{n_i} \right] - 3(N+1) \qquad [17.13]$$

Step 4. Given the level of significance ($\alpha$), determine the Kruskal-Wallis critical point ($\chi^2_{cp}$) as the upper (1–$\alpha$) × 100th percentage point from the chi-square distribution with ($k$–1) degrees of freedom (**Table 17-2** in **Appendix D**). Usually $\alpha$ is set equal to 0.05, so that the upper 95th percentage point of the chi-square distribution is needed.

Step 5. Compare the Kruskal-Wallis test statistic, *H*, against the critical point $\chi^2_{cp}$. If *H* is no greater than the critical point, conclude there is insufficient evidence of significant differences between any of the well group populations. If $H > \chi^2_{cp}$, however, conclude there is a significant difference between at least one pair of the well groups. Post-hoc comparisons are then necessary to determine whether any of the compliance wells significantly exceeds background (note that post-hoc comparisons are not necessary if using the Kruskal-Wallis test to merely identify spatial variability).

Step 6. In the case of a significant *H*-statistic that exceeds the critical point in Step 5, determine which compliance wells have elevated concentrations compared to background. This is done by comparing each compliance well against background, using a set of ***contrasts*** (as described for the parametric one-way ANOVA in **Section 17.1.1**).

To keep the test-wise or experiment-wise false positive rate close to the targeted (*i.e.*, nominal) rate of 5%, the individual contrasts should be set up as follows: Given (*k*–1) separate background-compliance contrasts, if (*k*–1) ≤ 5, run each contrast at a significance level equal to $\alpha^* = .05/(k–1)$. However, if (*k*–1) > 5, run each contrast at a significance level equal to $\alpha^* = .01$. Note that when there are more than 5 downgradient wells, this last provision will tend to raise the overall false positive rate above 5%.

Step 7. Since the background data is the first well group, the number of background observations is equal to $n_1$. For each of the remaining (*k*–1) well groups, compute the approximate rank-based standard error of the difference between each compliance well and background using equation [17.14]:

$$SE_i = \sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_1}+\frac{1}{n_i}\right)} \qquad [17.14]$$

Step 8. Let the average background rank be identified as $\overline{R}_b$. Compute the post-hoc *Z*-statistic for each of the (*k*–1) compliance wells for *i* = 2 to *k*, dividing the standard error in step 7 into the difference between the average rank at the compliance well and the background rank average, as shown below:

$$Z_i = \left(\overline{R}_i - \overline{R}_b\right)/SE_i \qquad [17.15]$$

Step 9. The *Z*-statistic in equation [17.15] has an approximate standard normal distribution under the null hypothesis that the *i*th compliance well is identical in distribution to background. The critical point ($z_{cp}$) can be found as the upper (1–α) × 100th percentage point of the normal distribution in **Table 10-1** of **Appendix D**.

Step 10. Compare the post-hoc *Z*-statistics for each of the (*k*–1) compliance wells against the critical point ($z_{cp}$). Any *Z*-statistic that exceeds the critical point provides significant evidence of an elevation over background in that compliance well at the α level of significance.

## ►EXAMPLE 17-2

Use the non-parametric Kruskal-Wallis test on the following data to determine whether there is evidence of possible toluene contamination at a significance level of $\alpha = 0.05$.

| | Toluene Concentration (ppb) | | | | |
|---|---|---|---|---|---|
| | Background Wells | | Compliance Wells | | |
| Month | Well 1 | Well 2 | Well 3 | Well 4 | Well 5 |
| 1 | <5 | <5 | <5 | <5 | <5 |
| 2 | 7.5 | <5 | 12.5 | 13.7 | 20.1 |
| 3 | <5 | <5 | 8.0 | 15.3 | 35.0 |
| 4 | <5 | <5 | <5 | 20.2 | 28.2 |
| 5 | 6.4 | <5 | 11.2 | 25.1 | 19.0 |

### SOLUTION

Step 1. Since non-detects account for 48% of these data, it would be very difficult to verify the assumptions of normality and equal variance necessary for a parametric ANOVA. Use the Kruskal-Wallis test instead, pooling both background wells into one group and treating each compliance well as a separate group. Note that $N = 25$ and $k = 4$.

Compute ranks for all the data including tied observations (*e.g.*, non-detects) as in the following table. Note that each non-detect is given the same midrank, equal to the average of the first 12 unique ranks.

| | Toluene Ranks | | | | |
|---|---|---|---|---|---|
| | Background Wells | | Compliance Wells | | |
| Month | Well 1 | Well 2 | Well 3 | Well 4 | Well 5 |
| 1 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 |
| 2 | 14 | 6.5 | 17 | 18 | 21 |
| 3 | 6.5 | 6.5 | 15 | 19 | 25 |
| 4 | 6.5 | 6.5 | 6.5 | 22 | 24 |
| 5 | 13 | 6.5 | 16 | 23 | 20 |
| Group Size | $n_1 = 10$ | | $n_2 = 5$ | $n_3 = 5$ | $n_4 = 5$ |
| Rank Sum | $R_{1\bullet} = 79$ | | $R_{2\bullet} = 61$ | $R_{3\bullet} = 88.5$ | $R_{4\bullet} = 96.5$ |
| Rank Mean | $\bar{R}_{1\bullet} = 7.9$ | | $\bar{R}_{2\bullet} = 12.2$ | $\bar{R}_{3\bullet} = 17.7$ | $\bar{R}_{4\bullet} = 19.3$ |

Step 2. Calculate the sum and average of the ranks in each group using equations [17.11] and [17.12]. These results are given in the above table.

Step 3. Compute the Kruskal-Wallis statistic $H$ using equation [17.13]:

$$H = \frac{12}{25 \cdot 26}\left[\frac{79^2}{10} + \frac{61^2}{5} + \frac{88.5^2}{5} + \frac{96.5^2}{5}\right] - (3 \cdot 26) = 10.56$$

Also compute the adjustment for ties with equation [17.10]. There is only one group of distinct tied observations — the non-detects — containing 12 samples. Thus, the adjusted Kruskal-Wallis statistic is given by:

$$H^* = 10.56 \Big/ \left[ 1 - \left( \frac{12^3 - 12}{25^3 - 25} \right) \right] = 11.87$$

Step 4. Determine the critical point of the Kruskal-Wallis test: with $\alpha = .05$, the upper 95th percentage point of the chi-square distribution with ($k$–1) = 4–1 = 3 degrees of freedom [$df$] is needed. **Table 17-2** of **Appendix D** gives $\chi^2_{cp} = \chi^2_{.95,3} = 7.81$.

Step 5. Since the observed Kruskal-Wallis statistic of 11.87 is greater than the chi-square critical point, there is evidence of significant differences between the well groups. Therefore, post-hoc pairwise comparisons are necessary.

Step 6. To determine the significance level appropriate for post-hoc comparisons, note there are three compliance wells that need to be tested against background. Therefore, each of these contrasts should be run at the $\alpha^* = 0.05/3 = 0.0167$ significance level.

Step 7. Calculate the standard error of the difference for the three contrasts using equation [17.14]. Since the sample size at each compliance well is five, the *SE* will be identical for each comparison, namely,

$$SE_i = \sqrt{\frac{25 \cdot 26}{12} \left( \frac{1}{10} + \frac{1}{5} \right)} = 4.031$$

Step 8. Form the post-hoc Z-statistic for each contrast using equation [17.15]:

$$\text{Well 3: } Z_2 = \left(12.2 - 7.9\right) \Big/ 4.031 = 1.07$$
$$\text{Well 4: } Z_3 = \left(17.7 - 7.9\right) \Big/ 4.031 = 2.43$$
$$\text{Well 5: } Z_4 = \left(19.3 - 7.9\right) \Big/ 4.031 = 2.83$$

Step 9. Find the upper $(1-\alpha^*) \times 100$th percentage point from the standard normal distribution in **Table 10-1** in **Appendix D**. With $\alpha^* = .0167$, this gives a critical point (by linear interpolation) of $z_{cp} = z_{.9833} = 2.127$.

Step 10. Since the *Z*-statistics at wells 4 and 5 exceed the critical point, there is significant evidence of increased concentration levels at wells 4 and 5, but not at well 3. ◄

## 17.2 TOLERANCE LIMITS

A *tolerance interval* is a concentration range designed to contain a pre-specified proportion of the underlying population from which the statistical sample is drawn (***e.g.***, 95 percent of all possible population measurements). Since the interval is constructed from random sample data, a tolerance interval is expected to contain the specified population proportion only with a certain level of statistical

confidence. Two coefficients are thus associated with any tolerance interval. One is the population proportion that the interval is supposed to contain, called the coverage ($\gamma$). The second is the degree of confidence with which the interval reaches the specified coverage. This is sometimes known as the tolerance coefficient or more simply, the confidence level ($1-\alpha$). A tolerance interval with 95% coverage and a tolerance coefficient of 90 percent is constructed to contain, on average, 95% of the distribution of all possible population measurements with a confidence probability of 90%.

A *tolerance limit* is a one-sided tolerance interval. The upper limit is typically of most interest in groundwater monitoring. Tolerance limits are a standard statistical method that can be useful in groundwater data analysis, especially as an alternative to *t*-tests or ANOVA for interwell testing. The RCRA regulations allow greater flexibility in the choice of $\alpha$ when using tolerance and prediction limits and control charts, so a larger variety of data configurations may be amenable to one of these approaches. The Unified Guidance still recommends prediction limits or control charts over tolerance limits for formal compliance testing in detection monitoring, and confidence intervals over tolerance limits in compliance/assessment monitoring when a background standard is needed.

An interwell tolerance limit constructed on background data is designed to cover all but a small percentage of the background population measurements. Hence background observations should rarely exceed the upper tolerance limit. By the same token, when testing a null hypothesis ($H_0$) that the compliance point population is identical to background, *compliance point* measurements also should rarely exceed the upper tolerance limit, unless $H_0$ is false. The upper tolerance limit thus gauges whether or not concentration measurements sampled from compliance point wells are too extreme relative to background.

## 17.2.1 PARAMETRIC TOLERANCE LIMITS

### BACKGROUND AND PURPOSE

To test the null hypothesis ($H_0$) that a compliance point population is identical to that of background, an upper tolerance limit with high coverage ($\gamma$) can be constructed on the sample background data. Coverage of 95% is usually recommended. In this case, random observations from a distribution identical to background should exceed the upper tolerance limit less than 5% of the time. Similarly, a tolerance coefficient or confidence level of at least 95% is recommended. This gives 95% confidence that the (upper) tolerance limit will contain at least 95% of the distribution of observations in background or in any distribution similar to background. Note that a tolerance coefficient of 95% corresponds to choosing a significance level ($\alpha$) equal to 5%. Hence, as with a one-way ANOVA, the overall false positive rate for a tolerance interval is set to approximately 5%.

Once the limit is constructed on background, each compliance point observation (perhaps from several different wells) is compared to the upper tolerance limit. This is different from the comparison of sample means in an ANOVA test. If any compliance point measurement exceeds the limit, the well from which it was drawn is flagged as showing a significant increase over background. Note that the factors $\kappa$ used to adjust the width of the tolerance interval (**Table 17-3** in **Appendix D**) are designed to provide *at least* 95% coverage of the parent population. Applied over many data sets, the *average* coverage of these intervals will often be close to 98% or more (see Guttman, 1970). Therefore, it would be unusual to find

more than 2 or 3 samples out of every 100 exceeding the tolerance limit under the null hypothesis. This fits with the purpose behind the use of a tolerance interval, which is to establish an upper limit on background that will rarely be exceeded, unless some change in the groundwater causes concentration levels to rise significantly at one or more compliance points.

Testing a large number of compliance point samples against such a background tolerance limit even under conditions of no releases practically ensures a few measurements will occasionally exceed the limit. The Unified Guidance therefore recommends that tolerance limits be used in conjunction with verification resampling of those wells suspected of possible contamination, in order to either verify or disconfirm the initial round of sampling and to avoid false positive results.

## REQUIREMENTS AND ASSUMPTIONS

Standard parametric tolerance limits assume normality of the sample background data used to construct the limit. This assumption is critical to the statistical validity of the method, since a tolerance limit with high coverage can be viewed as an estimate of a *quantile* or *percentile* associated with the *tail probability* of the underlying distribution. If the background sample is non-normal, a normalizing transformation should be sought. If a suitable transformation is found, the limit should be constructed on the transformed measurements and can then be *back-transformed* to the raw concentration scale prior to comparison against individual compliance point values.

If no transformation will work, a non-parametric tolerance limit should be considered instead. Unfortunately, non-parametric tolerance limits generally require a much larger number of observations to provide the same levels of coverage and confidence as a parametric limit. It is recommended that a parametric model be fit to the data if at all possible.

A tolerance limit can be computed with as few as three observations from background. However, doing so results in a high upper tolerance limit with limited statistical power for detecting increases over background. Usually, a background sample size of at least eight measurements will be needed to generate an adequate tolerance limit. If multiple background wells are screened in equivalent hydrostratigraphic positions and the data can reasonably be combined (**Chapter 5**), one should consider using pooled background data from multiple wells to increase the background sample size.

Like many tests described in the Unified Guidance, tolerance limits as applied to groundwater monitoring assume *stationarity* of the well field populations both temporally (*i.e.*, over time) and spatially. The data also needs to be statistically *independent*. Since an adequately-sized background sample will have to be amassed over time (in part to maintain enough temporal spacing between observations so that independence can be assumed), the background data should be checked for apparent *trends* or *seasonal effects*. As long the background mean is stable over time, the amassed data from a longer span of sampling will provide a better statistical description of the underlying background population.

As a primarily interwell technique, tolerance limits should only be utilized when there is minimal *spatial variability*. Explicit checks for spatial variation should be conducted using box plots and/or ANOVA.

In the usual test setting, one new compliance point observation from each distinct well is compared against the tolerance limit during each statistical evaluation. Under the null hypothesis of identical

populations, the compliance point measurements are assumed to follow the same distribution as background. Further, the compliance data are assumed to be mutually statistically independent. Such assumptions are almost impossible to check with only one new value per compliance well. However, periodic checks of the key assumptions are recommended after accumulating several sampling rounds of compliance data.

## PROCEDURE

Step 1. Calculate the mean $\bar{x}$, and the standard deviation *s*, from the background sample.

Step 2. Construct the one-sided upper tolerance limit as

$$TL = \bar{x} + \kappa\left(n, \gamma, 1-\alpha\right) \cdot s \qquad [17.16]$$

where $\kappa(n,\gamma,1-\alpha)$ is the one-sided normal tolerance factor found in **Table 17-3** of **Appendix D** associated with a sample size of *n*, coverage coefficient of γ, and confidence level of (1−α).

Equation [17.16] applies to normal data. If a transformation is needed to normalize the sample, the tolerance limit needs to be constructed on the transformed measurements and the limit back-transformed to the original concentration scale. If the limit was constructed, for example, on the logarithms of the original observations, where $\bar{y}$ and $s_y$ are the log-mean and log-standard deviation, the tolerance limit can be back-transformed to the concentration scale by exponentiating the limit. The tolerance limit is computed as:

$$TL = \exp\left[\bar{y} + \kappa\left(n, \gamma, 1-\alpha\right) \cdot s_y\right] \qquad [17.17]$$

Step 3. Compare each observation from the compliance well(s) to the upper tolerance limit found in Step 2. If any observation exceeds the tolerance limit, there is statistically significant evidence that the compliance well concentrations are elevated above background. Verification resampling should be conducted to verify or disconfirm the initial result.

## ►EXAMPLE 17-3

The table below consists of chrysene concentration data (ppb) found in water samples obtained from two background wells (Wells 1 and 2) and three compliance wells (Wells 3, 4, and 5). Compute the upper tolerance limit on background for coverage of 95% with 95% confidence and determine whether there is evidence of possible contamination at any of the compliance wells.

| | **Chrysene Concentration (ppb)** | | | | |
|---|---|---|---|---|---|
| Month | Well 1 | Well 2 | Well 3 | Well 4 | Well 5 |
| 1 | 19.7 | 10.2 | 68.0 | 26.8 | 47.0 |
| 2 | 39.2 | 7.2 | 48.9 | 17.7 | 30.5 |
| 3 | 7.8 | 16.1 | 30.1 | 31.9 | 15.0 |
| 4 | 12.8 | 5.7 | 38.1 | 22.2 | 23.4 |
| Mean | 19.88 | 9.80 | 46.28 | 24.65 | 28.98 |
| SD | 13.78 | 4.60 | 16.40 | 6.10 | 13.58 |

SOLUTION

Step 1. A Shapiro-Wilk test of normality on the pooled set of eight background measurements gives $W = 0.7978$ on the original scale and $W = 0.9560$ after log-transforming the data, suggesting that the data are better fit by a lognormal distribution. Therefore, construct the tolerance limit on the logged observations, listed below along with the log-means and log-standard deviations.

| | Log Chrysene log(ppb) | | | | |
|---|---|---|---|---|---|
| Month | Well 1 | Well 2 | Well 3 | Well 4 | Well 5 |
| 1 | 2.981 | 2.322 | 4.220 | 3.288 | 3.850 |
| 2 | 3.669 | 1.974 | 3.890 | 2.874 | 3.418 |
| 3 | 2.054 | 2.779 | 3.405 | 3.463 | 2.708 |
| 4 | 2.549 | 1.740 | 3.640 | 3.100 | 3.153 |
| Mean | 2.813 | 2.204 | 3.789 | 3.181 | 3.282 |
| SD | 0.685 | 0.452 | 0.349 | 0.253 | 0.479 |
| BG Mean | 2.509 | | | | |
| BG SD | 0.628 | | | | |

Step 2. Compute the upper tolerance limit on the pooled background data using the logged chrysene concentration data. The tolerance factor for a one-sided upper normal tolerance limit with 95% coverage and 95% probability and $n = 8$ observations is equal to (from **Table 17-3** of **Appendix D**) $\kappa = 3.187$. Therefore, the upper tolerance limit is computed using equation [17.17] as:

$$TL = \exp\left[2.509 + 3.187 \times 0.628\right] = 90.96 \text{ ppb}$$

Step 3. Compare the measurements at each compliance well to the upper tolerance limit, that is $TL$ = 90.96 ppb. Since none of the original chrysene concentrations exceeds the upper $TL$, there is insufficient evidence of chrysene contamination in these data. ◄

## 17.2.2 NON-PARAMETRIC TOLERANCE INTERVALS

### BACKGROUND AND PURPOSE

When an assumption of normality cannot be justified especially with a significant portion of non-detect observations, the use of non-parametric tolerance intervals should be considered. The upper tolerance limit in a non-parametric setting is usually chosen as an order statistic of the sample data (Guttman, 1970), commonly the maximum value or maybe the second or third largest value observed.

Because the maximum observed background value is often taken as the upper tolerance limit, non-parametric tolerance intervals are easy to construct and use. The sample data needs to be ordered, but no

ranks need be assigned to the concentration values other than to determine the largest measurements. This also means that non-detects do not have to be uniquely ordered or handled in any special manner.

One advantage to using a maximum concentration instead of assigning ranks to the data (Wilcoxon rank-sum or Kruskal-Wallis tests) is that non-parametric tolerance intervals are reflective of actual concentration magnitudes. Another advantage is that unless all the background data are non-detect, the maximum value will be a detected concentration leading to a well-defined upper tolerance limit. If all the sample data are non-detect, an RL (*e.g.*, the lowest achievable *quantitation limit* [QL]) may serve as an approximate upper tolerance limit.

## REQUIREMENTS AND ASSUMPTIONS

Unlike parametric tolerance intervals, the desired coverage ($\gamma$) or confidence level ($1-\alpha$) cannot be pre-specified using a non-parametric limit. Instead, the *achieved* coverage and/or confidence level depends entirely on the background sample size ($n$) and the order statistic chosen as the upper tolerance limit (*e.g.*, the maximum value). Guttman (1970) has shown that the *coverage* of the limit follows a *beta probability density* with cumulative distribution:

$$I_t\left(n-m+1,m\right)=\int_{u=0}^{t}\frac{\Gamma\left(n+1\right)}{\Gamma\left(n-m+1\right)\Gamma\left(m\right)}u^{n-m}\left(1-u\right)^{m-1}du \qquad [17.18]$$

where $n$ = sample size and $m$ = [($n$+1)–(rank of upper tolerance limit value)]. If the background maximum is selected as the tolerance limit, its rank is equal to $n$ and so $m$ = 1. If the second largest value is chosen as the limit, its rank would be equal to ($n$–1) giving $m$ = 2.

As a non-parametric procedure, no distributional model must be fit to the background measurements. It is assumed, however, that the compliance point data follow the same distribution as background — even if unknown — under the null hypothesis. Even though no distributional model is assumed, order statistics of any random sample follow certain probability laws as noted above. Since the beta distribution is closely related to the more familiar *binomial distribution*, Guttman showed that in order to construct a non-parametric tolerance interval with at least $\gamma$ coverage and ($1-\alpha$) confidence probability, the number of (background) samples should be chosen such that:

$$\sum_{t=m}^{n}\binom{n}{t}\left(1-\gamma\right)^{t}\gamma^{n-t}\geq 1-\alpha \qquad [17.19]$$

If the background maximum is selected as the upper tolerance limit, so that $m$ = 1, this inequality reduces to the simpler form

$$1-\gamma^{n}\geq 1-\alpha\,. \qquad [17.20]$$

**Table 17-4** in **Appendix D** provides minimum coverage levels with 95% confidence for various choices of $n$, using either the maximum sample value or the second largest measurement as the tolerance limit. As an example, with $n$ = 16 background measurements, the minimum coverage is $\gamma$ = 83% if the background maximum is designated as the upper tolerance limit and $\gamma$ = 74% if the tolerance limit is

taken as the second largest background value. In general, **Table 17-4** of **Appendix D** illustrates that if the underlying distribution is unknown, *more background samples are needed compared to the parametric setting in order to construct a tolerance interval with sufficiently high coverage.* Parametric tolerance intervals do not require as many background samples precisely because the form of the underlying distribution is assumed to be known.

An alternate way to construct an appropriate tolerance limit is to calculate the maximum confidence level for various choices of $n$ guaranteeing at least 95% coverage. With $n = 8$ background measurements, the approximate confidence level is at most 34% when the largest value is taken as the tolerance limit and only 6% if the second largest value is taken as the tolerance limit. Clearly, it is advantageous to fit a parametric distributional model to the data if at all possible unless $n$ is fairly large.

Although non-parametric tolerance limits do not require an assumption of normality, other assumptions of tolerance limits apply equally to the parametric and non-parametric versions. Specifically, the sample data should be statistically *independent* and show no evidence of *autocorrelation*, *trends*, or *seasonal effects* in background. Applied as an interwell test, there should also be minimal to no natural on-site *spatial variation*.

By construction, outliers in background can be a particular problem for non-parametric tolerance limits, especially if the background maximum is chosen as the upper limit. A limit based on a large, extreme outlier will result in a test having little power to detect increases in compliance wells. Consequently, the background sample should be screened ahead of time for possible *outliers* (**Chapter 12**). Confirmed outliers should be removed from the data set before setting the tolerance limit.

An important caveat to this advice is that almost all statistical outlier tests depend crucially on the ability to fit the remaining data (minus the suspected outlier(s)) to a known statistical distribution. In those cases where a non-parametric tolerance limit is selected because of a large fraction of non-detects, fitting the data to a distributional model may be difficult or impossible, negating formal outlier tests. As an alternative, the non-parametric upper tolerance limit could be set to a different order statistic in background (*i.e.*, other than the maximum), to provide some insurance against possible large outliers. This strategy will work provided there are enough background measurements to allow for adequately high coverage and confidence in the resulting limit.

## PROCEDURE

Step 1. Sort the set of background data into ascending order and choose either the largest or second largest measurement as the upper *TL*. Use **Table 17-4** in **Appendix D** to determine the coverage $\gamma$ associated with 95% or 99% confidence. Note also that if the largest or second largest measurement is a non-detect, the upper tolerance limit should be set to the RL most appropriate to the data (*e.g.*, the lowest achievable practicable quantification limit [PQL]).

Step 2. Compare each compliance point measurement against the upper tolerance limit. Identify significant evidence of possible contamination at any compliance well in which one or more measurements exceed the upper tolerance limit. If the upper tolerance limit equals the RL, a violation should be flagged anytime a detected value is quantified above the RL.

Step 3. Because the risk of false positive errors is greatly increased if either the confidence level or coverage drop substantially below 95%, both of these parameters should be routinely reported

and noted as being below the target levels. One should also strongly consider comparing one or more verification resamples against the upper tolerance limit before identifying a clear violation.

►EXAMPLE 17-4

Use the following copper background data to establish a non-parametric upper tolerance limit and determine if either compliance well shows evidence of copper contamination.

| | **Copper Concentration (ppb)** | | | | |
|---|---|---|---|---|---|
| | Background Wells | | | Compliance Wells | |
| Month | Well 1 | Well 2 | Well 3 | Well 4 | Well 5 |
| 1 | <5 | 9.2 | <5 | | |
| 2 | <5 | <5 | 5.4 | | |
| 3 | 7.5 | <5 | 6.7 | | |
| 4 | <5 | 6.1 | <5 | | |
| 5 | <5 | 8.0 | <5 | 6.2 | <5 |
| 6 | <5 | 5.9 | <5 | <5 | <5 |
| 7 | 6.4 | <5 | <5 | 7.8 | 5.6 |
| 8 | 6.0 | <5 | <5 | 10.4 | <5 |

SOLUTION

Step 1. The pooled background data in Wells 1, 2, and 3 have a maximum observed value of 9.2 ppb. Set the 95% confidence upper tolerance limit equal to this value. Because 24 background samples are available, **Table 17-4** in **Appendix D** indicates that the minimum coverage is equal to 88%. To increase either the coverage, more background samples would have to be collected.

Step 2. Compare each sample in compliance Wells 4 and 5 to the upper tolerance limit. Since none of the measurements at Well 5 is above 9.2 ppb, while one sample from Well 4 is above the limit, conclude that there may be significant evidence of copper contamination at Well 4 but not Well 5.

Step 3. Note that with only 88% coverage and 24 background samples, the risk of a false positive result is more than 10%. Well 4 should be resampled to determine whether the exceedance is replicated. ◄

## 17.3 TREND TESTS

The Unified Guidance recommends *trend testing* as an intrawell alternative to prediction limits or control charts when those methods are not suitable. Prediction limits and control charts (as well as *t*-tests and ANOVA) all involve a comparison of compliance and background populations under the key assumption that the underlying concentration distributions are *stationary over time*. That is, the populations are presumed to have stable (*i.e.*, roughly constant) means over the period of sampling prior to statistical evaluation.

Unfortunately, there is no guarantee that groundwater populations will remain stable during long-term monitoring. Because sampling at many sites is generally done on a quarterly, semi-annual, or annual basis, it will generally take one to two years or more to collect enough background data to run the statistical tests discussed in the Unified Guidance. Over this length of time, the statistical characteristics of groundwater may or may not change in significant ways.

If background groundwater conditions are in a state of flux, trend tests provide a significant advantage over both intrawell prediction limits and control charts. Both of the latter methods involve a designation of some portion of the historical sampling record as the intrawell background for a given compliance well. Ideally, this intrawell background should consist of measurements known to be uncontaminated and which represent a random sample from a stable underlying population, just as with *t*-tests and ANOVA. If the mean and/or standard deviation of the underlying population ***changes*** while intrawell background is being compiled, results of either prediction limit or control chart tests against more recently collected data can be severely biased or altogether inaccurate.

One drawback to the Shewhart-CUSUM control charts presented in **Chapter 20** is that they are somewhat sensitive to the parametric assumption of underlying normality. If the measurements are lognormal rather than normal, for instance, the nominal performance characteristics (***i.e.***, Type I error rate and statistical power) of control charts are significantly affected. By the same token, control charts are impacted if the intrawell background contains a large fraction of non-detects. Non-detect adjustments can sometimes be made to the baseline data via methods discussed in **Chapter 15**, but if a normalizing transformation or adjustment is not successful, no straightforward non-parametric control chart exists.

Consequently, neither prediction limits nor control charts are appropriate for every circumstance where an intrawell comparison may be warranted or necessary. Thus, ***the Unified Guidance recommends that users consider trend testing as an alternative to prediction limits or control charts when those methods are not suitable as intrawell techniques***. Tests for trend are specifically designed to identify those groundwater populations whose mean concentrations are not stationary over time, but rather are increasing (or decreasing) by measurable amounts. Ultimately, the goal of any reasonable detection or compliance/assessment monitoring program is to determine whether or not the concentration levels of key contaminants or indicator parameters have significantly increased during the period of monitoring and, if so, whether the increase is attributable to facility waste management practices.

The detection of trends is a complex subject. Whole textbooks are devoted to the more general topic of *time series analysis*, including the identification and modeling of time trends — step functions, linear and quadratic trends, exponential growth, *etc.* The Unified Guidance only attempts to identify the simplest kind of linear increases, not the specification or testing of more complex models. The methods described below are all designed to effectively test for (increasing) linear trends, though they will also identify simple increases over time when a trend is present but does not follow a strictly linear pattern.

The Unified Guidance recommends using trend tests in detection monitoring to measure the extent and nature of an apparent concentration increase, especially to determine whether or not the increase occurs consistently over time. Two questions are of particular interest: 1) is there a statistically significant, (positive) trend over the period of monitoring? and 2) what is the nature (***i.e.***, slope and intercept) of the trend? By identifying a positive trend, one can show that contaminant levels have gotten worse compared to early measurements from the well being tested. Furthermore, by measuring the nature

of the trend, including the average rate of increase per unit of time, one can estimate how rapidly concentration levels are increasing and the current mean- or median-level magnitude of contamination. Such information can provide an invaluable portrait of the changes occurring on-site and probably offers the most compelling evidence — under these conditions — for demonstrating that the basic null hypothesis of detection monitoring has been violated.

## 17.3.1 LINEAR REGRESSION

### BACKGROUND AND PURPOSE

The most common way to measure a linear trend is to compute a *linear regression* of concentration data when plotted against the time or date of sample collection. By way of interpretation, each point along a linear regression trend line is an estimate of the true mean concentration *at that point in time*. Thus, a linear regression can be used to assess whether or not the population mean at a compliance well has significantly increased or decreased.

Linear regression is a standard technique in statistics textbooks and many data analysis software packages. It is more generally applicable to linear relationships between any pair of random variables and not simply to time trends. Good references for performing linear regression and for checking and verifying its assumptions include Draper and Smith (1998) and Cook and Weisberg (1999).

Unlike prediction limits or control charts which are constructed using only the background data, trend tests including linear regression are computed with all available earlier and more recent data at the compliance well of interest. One then might incorrectly assume that a comparison against intrawell background is not being conducted. But an intrawell comparison does occur with a trend test. Statistical identification of a structured pattern of increase from the first portion of the sampling record to more recent data indicates that concentration levels are no longer similar to intrawell background, but have risen more than expected by chance.

Statistical identification of a positive trend involves testing the estimated slope coefficient from the linear regression trend line. A specially constructed *t*-test is used to make this determination, as described below. If this test is significant, the slope is judged to be different from zero, indicating that a change in concentration levels has occurred over the period of sampling represented by the data set.

### REQUIREMENTS AND ASSUMPTIONS

Linear regression as a parametric statistical technique makes a number of underlying assumptions. Among the most important of these are that the regression residuals (*i.e.*, the difference between each concentration measurement and its predicted value from the regression equation) are approximately normal in distribution, homoscedastic (*i.e.*, equal in variance at different times and for different mean concentration levels), and statistically independent. Significant skewness or the presence of outliers can bias or invalidate the results of a trend test based on linear regression. Furthermore, standard linear regression methods do not account for non-detects or missing data values at selected sampling events.

Because the key assumptions for linear regression depend not on the original measurements but rather on the regression residuals, a tentative trend line needs to first be constructed before its

assumptions can be checked. Once a linear regression on time is fitted to the data, the residuals around the trend line need to be computed and then tested for normality, apparent skewness, and equal variance over time. This last assumption is particularly important to testing whether the slope of an apparent trend is statistically different from zero (a zero slope indicating that well concentrations have not changed over time).

Inferences around a linear regression are generally appropriate when three conditions hold: 1) the residuals from the regression are approximately normal or at least reasonably symmetric in distribution; 2) a scatter plot of residuals versus concentrations indicates a scatter cloud of essentially uniform *vertical thickness or width* (i.e., the scatter cloud does not tend to increase in width with the level of concentration which would suggest a proportional effect between the underlying population mean and variance); and 3) a scatter plot of residuals versus time also exhibits a uniformly thick scatter cloud. If the thickness or width is substantially different at distinct time points, the assumption of equal variances over time may not be true.

If any of these conditions is substantially violated, it may indicate that the basic trend is either non-linear or the magnitude of the variance is not independent of the mean concentration level and/or the time of sampling. One possible remedy is to try a transformation of the concentration data and re-estimate the linear regression. This will change the interpretation of the estimated regression from a linear trend of the form $y = a + bt$, where $y$ and $t$ represent concentration and time respectively, to a non-linear pattern. As an example, if the concentration data are log-transformed, the regression equation will have the form $\log y = a + bt$. Back-transformed to the original concentration scale, the trend function will then have the form $y = \exp\left(a + bt\right)$.

In transforming the regression data this way, the estimated trend in the concentration domain (after back-transforming) no longer represents the original mean. Rather, the transformation induces a ***bias*** when converted back to the raw-scale data. If a log transformation is used, for instance, the back-transformed trend line will represent the raw-scale *geometric mean* and not the *arithmetic mean*. As with Student's t-tests on lognormal data (**Chapter 16**), demonstrating that the *geometric* mean is increasing also implies that the *arithmetic* mean has risen so long as the regression residuals are homoscedastic.

A minimum of 8 to 10 measurements is generally necessary to compute a linear regression, especially to estimate the variance around the trend line (known as the *mean squared error* or MSE). The regression residuals should be statistically independent, an assumption that can be approximately verified via one of the autocorrelation tests of **Chapter 14**.

One last assumption is that there should be few if any non-detects when computing a linear regression. As a matter of common sense, a significant increasing or decreasing trend should be based on reliably quantified measurements. If this is not the case, the user should check to see whether the "trend" may be an artifact induced by changes in detection and/or quantitation limits over time. The concentration levels of a series of non-detects may appear to be decreasing, for instance, simply because analytical methods have improved over the years leading to lower RLs. Such artifacts of plotting and data reporting should not be considered real trends.

When the assumptions of linear regression cannot be verified at least approximately, a non-parametric trend method should be considered instead. **Sections 17.3.2** and **17.3.3** discuss the *Mann-*

*Kendall test* for trend and the *Theil-Sen trend line*. These methods can be particularly valuable when constructing trends on data sets containing non-detects.

## PROCEDURE

Step 1. Construct a time series plot of the compliance point measurements. If a discernible trend is evident, compute a linear regression of concentration against sampling date (time), letting $x_i$ denote the *i*th concentration value and $t_i$ denote the *i*th sampling date. Estimate the linear slope $\hat{b}$ with the formula:

$$\hat{b} = \sum_{i=1}^{n} (t_i - \bar{t}) \cdot x_i \Big/ (n-1) \cdot s_t^2 \qquad [17.21]$$

This estimate then leads to the regression equation, given by:

$$\hat{x}_t = \bar{x} + \hat{b} \cdot (t - \bar{t}) \qquad [17.22]$$

where $\bar{t}$ denotes the mean sampling date, $s_t^2$ is the variance of sampling dates, $\bar{x}$ is the mean concentration level, and $\hat{x}_t$ represents the estimated mean concentration at time *t*.

Note: though the variable *t* above represents time, it could just as easily signify another variable, perhaps a second constituent for which an association with *x* is estimated.

Step 2. Compute the regression residual at each sampling event *i* with equation [17.23]:

$$r_i = x_i - \hat{x}_i \qquad [17.23]$$

Check the set of residuals for lack of normality and significant skewness using the techniques in **Chapter 10**. Also, plot the residuals against the estimated regression values ($\hat{x}_i$) to check for non-uniform vertical thickness in the scatter cloud. Make a similar check by plotting the residuals against the sampling dates ($t_i$).

If the residuals are non-normal and substantially skewed and/or the scatter clouds appear to have a definite pattern (***e.g.***, funnel-shaped; "U"-shaped; or, residuals mostly positive on one end of graph and mostly negative on the other end, instead of randomly scattered around the horizontal line $r = 0$), repeat **Steps 1** and **2** after first attempting a normalizing transformation.

Step 3. Calculate the estimated variance around the regression line (also known as the *mean squared error* [MSE]) with equation [17.24]:

$$s_e^2 = \frac{1}{n-2} \sum_{i=1}^{n} r_i^2 \qquad [17.24]$$

Step 4. Compute the standard error of the linear regression slope coefficient using the $s^2_e$ result from Step 3 in equation [17.25]:

$$se(\hat{b}) = \sqrt{s^2{}_e \Big/ \sum_{i=1}^{n} (t - \bar{t})^2} \qquad [17.25]$$

Step 5. Test whether the trend is significantly different from zero by forming the *t*-statistic ratio in equation [17.26]:

$$t_b = \hat{b} / se(\hat{b}) \qquad [17.26]$$

This *t*-statistic ($t_b$) has $n$–2 degrees of freedom [*df*]. Given a level of significance (α), choose the critical point ($t_{cp}$) for the test as the (1– α) × 100th percentage point of the Student's *t*-distribution with ($n$–2) *df* or $t_{cp} = t_{1-\alpha,n-2}$. Compare $t_b$ against the critical point. If $t_b > t_{cp}$, conclude that the slope of the trend is both positive and significantly different from zero at the α-level of significance. If $t_b < -t_{cp}$, conclude there is a significant decreasing trend. If neither exists, there is insufficient evidence of an increasing or decreasing trend.

## ►EXAMPLE 17-5

The following groundwater chloride measurements ($n$ = 19) were collected over a five-year period at a solid waste landfill. Test for a significant trend at the α = 0.01 level using linear regression.

| Sample Date | Chloride (ppm) | Elapsed Days | Residuals |
|---|---|---|---|
| 2002-03-18 | 11.5 | 76 | –0.25 |
| 2002-05-14 | 12.6 | 133 | 0.67 |
| 2002-08-22 | 13.8 | 233 | 1.56 |
| 2003-02-12 | 12.3 | 407 | –0.48 |
| 2003-05-29 | 12.8 | 513 | –0.30 |
| 2003-08-18 | 13.2 | 594 | –0.15 |
| 2003-11-20 | 14.1 | 688 | 0.45 |
| 2004-02-19 | 13.3 | 779 | –0.63 |
| 2004-04-26 | 13.1 | 846 | –1.04 |
| 2004-07-29 | 13.2 | 940 | –1.23 |
| 2004-11-09 | 15.3 | 1043 | 0.56 |
| 2005-02-24 | 15.0 | 1150 | –0.08 |
| 2005-06-14 | 15.2 | 1260 | –0.22 |
| 2005-08-23 | 15.8 | 1330 | 0.17 |
| 2005-10-17 | 16.1 | 1385 | 0.30 |
| 2006-02-08 | 15.1 | 1499 | –1.06 |
| 2006-04-27 | 16.4 | 1577 | 0.00 |
| 2006-08-10 | 17.7 | 1682 | 0.98 |
| 2006-10-26 | 17.7 | 1759 | 0.74 |

### SOLUTION

Step 1. Check for an apparent trend on a time series plot (**Figure 17-2**). Since the chloride values are increasing in reasonably linear fashion, compute the tentative regression line using equations [17.21] and [17.22]. To compute the slope estimate, first convert the sample dates to elapsed days using a starting date prior to the first event. In this case, choose an arbitrary starting date of 2002-01-01 as zero and compute the elapsed days as listed in the table above.

Using elapsed days as the time variable, compute the sample mean and variance to get:

$$\bar{t} = 941.79 \text{ days}$$
$$s_t^2 = 279374.3 \text{ days}^2$$

Then compute the tentative slope as:

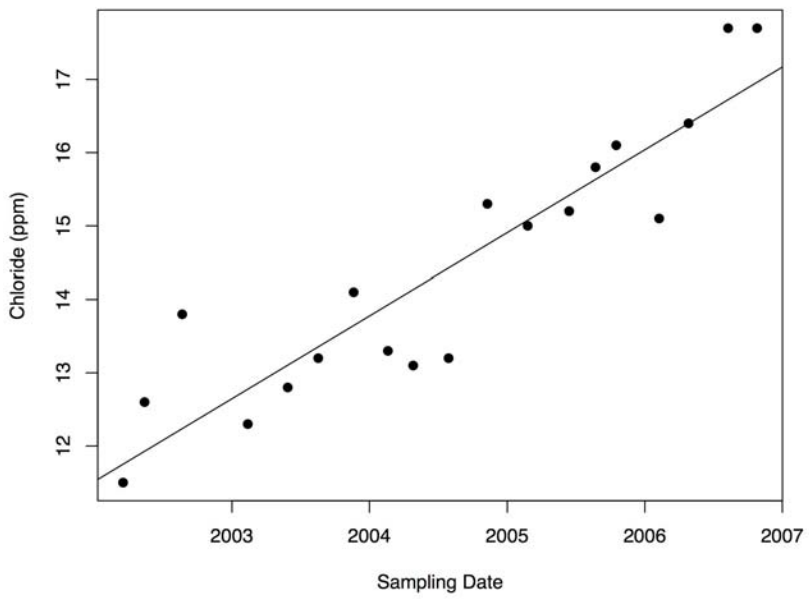
$$\hat{b} = \left[(76-941.79)\cdot 11.5 + \ldots + (1759-941.79)\cdot 17.7\right] / \left[(19-1)\cdot 279374.3\right] = .0031$$

and the regression line itself as:

$$\hat{x}_t = \bar{x} + \hat{b}\cdot(t-\bar{t}) = 14.432 + .0031\cdot(t-941.79)$$

where the mean chloride value is $\bar{x} = 14.432$ ppm. The regression line is overlaid on the scatter plot in **Figure 17-2**.

Figure 17-2. Time Series Plot of Chloride (ppm) Overlaid With Linear Regression




Step 2. Calculate the regression residual at each sampling event using equation [17.23]. This involves computing an estimated concentration along the regression line for each sampled time (*t*) and then subtracting from the observed concentration. For example, the residual at t = 407 is

$$x_t - \hat{x}_t = 12.3 - 12.78 = -0.48$$

All the residuals are listed in the table above. Then check the residuals for normality, homoscedasticity, and lack of association with the predicted values from the regression line.

**Figure 17-3** is a probability plot of the residuals, indicating good agreement with normality. **Figure 17-4** is a scatter plot of the residuals versus sampling date and **Figure 17-5** is a scatter plot of the residuals versus predicted values from the trend line. Both of these last plots do not exhibit any particular trends or patterns with sampling date or the trend line predicted values; the residuals are fairly randomly scattered.

Step 3. Compute the MSE of the regression using the squared residuals in equation [17.24] to get

$$s_e^2 = \frac{1}{n-2} \cdot \sum_{i=1}^{n} r_i^2 = \frac{1}{17} \cdot \left[ (-.25)^2 + (.67)^2 + \ldots + (.74)^2 \right] = 0.5628$$

Step 4. Calculate the standard error of the regression slope coefficient using equation [17.25]:

$$se\left(\hat{b}\right) = \sqrt{s^2{}_e \Big/ \sum_{i=1}^{n} (t - \bar{t})^2} = \sqrt{.5628 \Big/ \left[ (76 - 941.79)^2 + \ldots + (1759 - 941.79)^2 \right]} = .00033$$

Step 5. Form the *t*-statistic ratio with formula [17.26] to get:

$$t_b = \hat{b} \big/ se\left(\hat{b}\right) = 0.0031 / 0.00033 = 9.39$$

Since $\alpha = 0.01$, compare this value to a critical point equal to the 99th percentile of a Student's *t*-distribution with $(n-2) = 17$ degrees of freedom, that is, $t_{cp} = t_{.99,17} = 2.567$. Since the *t*-statistic is substantially larger than the critical point, conclude the upward trend is significant at the 1% α-level. ◄

Figure 17-3. Probability Plot of Chloride Regression Residuals

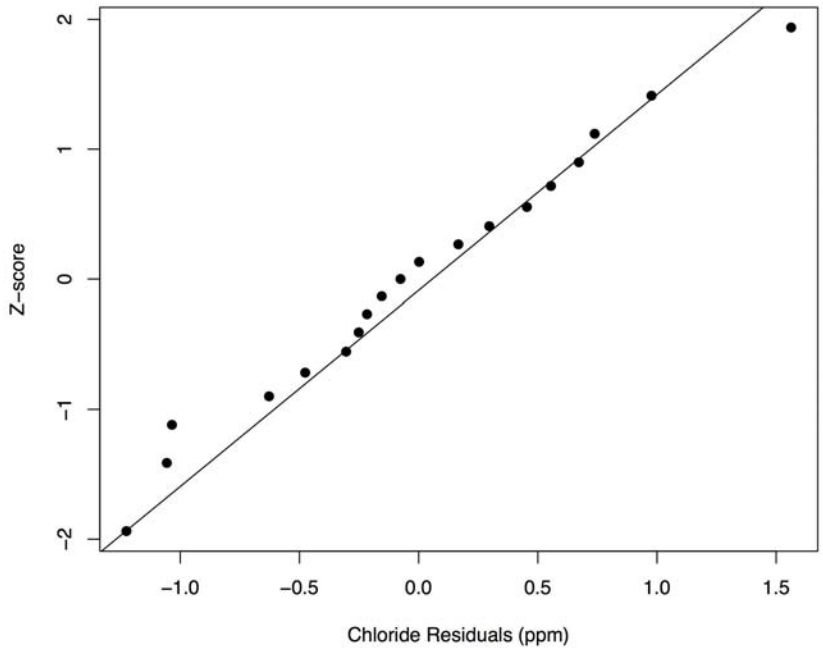



Figure 17-4. Scatter Plot of Chloride Residuals vs. Sampling Date

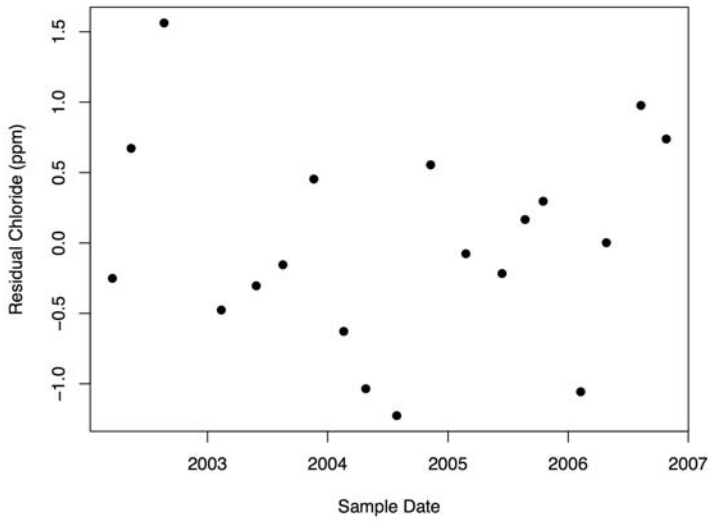
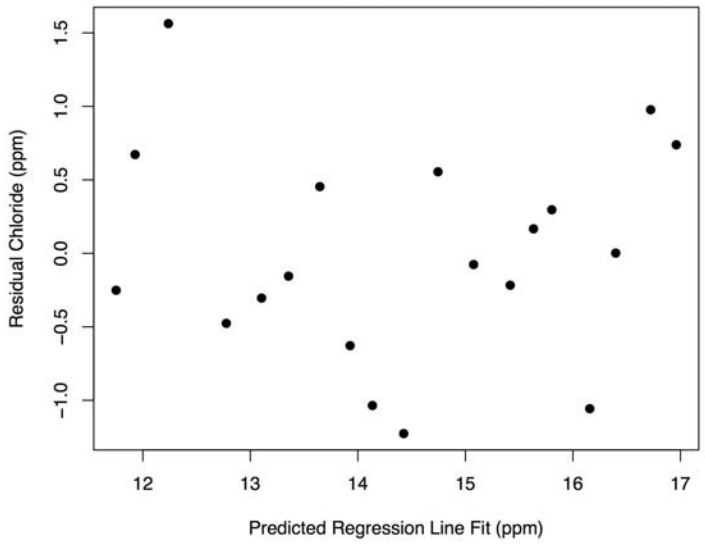


Figure 17-5. Scatter Plot of Chloride Residuals vs. Predicted Regression Fits




## 17.3.2 MANN-KENDALL TREND TEST

### BACKGROUND AND PURPOSE

The Mann-Kendall test (Gilbert, 1987) is a non-parametric test for linear trend, based on the idea that a lack of trend should correspond to a time series plot fluctuating randomly about a constant mean level, with no visually apparent upward or downward pattern. If an *increasing* trend really exists, the sample taken first from any randomly selected pair of measurements should on average have a *lower* concentration than the measurement collected at a later point. The Mann-Kendall statistic is computed by examining all possible pairs of measurements in the data set and scoring each pair as follows. An earlier measurement less in magnitude than a later one is assigned a value of 1. If an earlier value is greater in magnitude than a later sample, the pair is tallied as –1; two identical measurement values are assigned 0.

After scoring each pair in this way and adding up the total to get the Mann-Kendall statistic (*S*), a positive value of *S* implies that a majority of the differences between earlier and later measurements are positive, suggestive of an upward trend over time. Likewise, a negative value for *S* implies that a majority of the differences between earlier and later values are negative, suggestive of a decreasing trend. A value near zero indicates a roughly equal number of positive and negative differences. This would be expected if the measurements were randomly fluctuating about a constant mean with no apparent trend.

To account for randomness and inherent variability in the sample, the Mann-Kendall test is based on the critical ranges of the statistic *S* likely to occur under stationary conditions. The larger the absolute

value of *S*, the stronger the evidence for a real increasing or decreasing trend. The critical points for identifying a trend get larger as the level of significance ($\alpha$) drops. Only if the absolute value of the test statistic (*S*) is larger than the critical point is a statistically significant increasing or decreasing trend indicated.

## REQUIREMENTS AND ASSUMPTIONS

As a non-parametric procedure, the Mann-Kendall test does not require the underlying data to follow a specific distribution. Ranks of the data are not explicitly used in forming the test statistic as with the Wilcoxon rank-sum. Only the relative magnitudes of the concentration values are needed to compute *S*, not the actual concentrations themselves. Non-detects can be treated by assigning them a common value lower than any of the detected measurements. Any pair of tied values or any pair of non-detects is simply given a score of 0 in the calculation of the Mann-Kendall statistic *S*.

This treatment of non-detects is an imperfect remedy since it is usually impossible to know whether censored values are actually tied in magnitude. Further complications are introduced when there are multiple RLs and/or an intermingling of detected values and RLs. Lab qualifiers may be used to aid the scoring of pairs that involve non-detects or estimated concentrations. Instead of treating all non-detects as tied, consider 'undetected or U' values as the lowest in magnitude, other non-detects as higher in magnitude than U's but lower than estimated concentrations ('J' or 'E' values). In this way, a richer scoring of the sample pairs may be possible.

When the sample size *n* becomes large, exact critical values for the statistic *S* are not readily available. However, as a sum of identically-distributed random quantities, the behavior of *S* for larger *n* tends to approximate the normal distribution by the Central Limit Theorem. Therefore a normal approximation to *S* can be used for $n > 10$[1]. In this case, a standardized *Z*-statistic is formed by first computing the expected mean value and standard deviation of *S*. From the discussion above, when no trend is present, positive differences in randomly selected pairs of measurements should balance negative differences, so the expected mean value of *S* under the null hypothesis of no trend is simply zero. The standard deviation of *S* can be computed using equation [17.27]:

$$SD\left[S\right] = \sqrt{\frac{1}{18}\left[n(n-1)(2n+5) - \sum_{j=1}^{g} t_j\left(t_j-1\right)\left(2t_j+5\right)\right]} \qquad [17.27]$$

where *n* is the sample size, *g* represents the number of groups of ties in the data set (if any), and $t_j$ is the number of ties in the *j*th group of ties. If no ties or non-detects are present, equation [17.27] reduces to the simpler form:

$$SD\left[S\right] = \sqrt{\frac{1}{18}n(n-1)(2n+5)} \qquad [17.28]$$

[1] Guidance **Table 17-5** contains exact confidence levels up to $n = 10$. Exact confidence levels for $n \leq 20$ have been developed in (Hollander & Wolfe, 1999), Table A.30. These might be preferentially used if sample sizes are fairly small and the data contain non-detect values.

Once the standard deviation of *S* has been derived, the standardized *Z*-statistic for an increasing (or decreasing) trend is formed using the equation:

$$Z = \left(|S| - 1\right) \Big/ SD\left[S\right] \qquad [17.29]$$

Note that although the expected mean value of *S* is zero, applying the continuous normal to the discrete *S* distribution is an approximation. Therefore, a ***continuity correction*** is made to *Z* by first subtracting 1 from the absolute value of *S*. The final *Z*-statistic can then be compared to an $\alpha$-level critical point taken from **Table 10-1** in **Appendix D** to complete the test.

## PROCEDURE

Step 1. Order the data set by sampling event or time of collection, $x_1$, $x_2$, to $x_n$. Then consider all possible differences between distinct pairs of measurements, $(x_j - x_i)$ for $j > i$. For each pair, compute the ***sign*** of the difference, defined by:

$$\operatorname{sgn}\left(x_j - x_i\right) = \begin{cases} 1 & \text{if } \left(x_j - x_i\right) > 0 \\ 0 & \text{if } \left(x_j - x_i\right) = 0 \\ -1 & \text{if } \left(x_j - x_i\right) < 0 \end{cases} \qquad [17.30]$$

Pairs of tied values including non-detects, will receive scores of zero using equation [17.30].

Step 2. Compute the Mann-Kendall statistic *S* using equation [17.31]:

$$S = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \operatorname{sgn}\left(x_j - x_i\right) \qquad [17.31]$$

In equation [17.31] the summation starts with a comparison of the very first sampling event against each of the subsequent measurements. Then the second event is compared with each of the samples taken after it (*i.e.*, the third, fourth, fifth, *etc.*). Following this pattern is probably the most convenient way to ensure that all distinct pairs are tallied in forming *S*. For a sample of size *n*, there will be $n \cdot (n\text{-}1)/2$ distinct pairs.

Step 3. If $n \leq 10$, and given the level of significance ($\alpha$), determine the critical point $s_{cp}$ from **Table 17-5 of Appendix D.** If $S > 0$ and $|S| > s_{cp}$, conclude there is statistically significant evidence of an increasing trend at the $\alpha$ significance level. If $S < 0$ and $|S| > s_{cp}$, conclude there is statistically significant evidence of a decreasing trend. If $|S| \leq s_{cp}$, conclude there is insufficient evidence to identify a significant trend.

Step 4. If $n > 10$, determine the number of groups of ties ($g$) and the number of tied values in each group of ties ($t_j$). Then use equation [17.27] to compute the standard deviation of *S* and equation [17.29] in turn to compute the standardized *Z*-statistic.

Step 5. Given the significance level ($\alpha$), determine the critical point $z_{cp}$ from the standard normal distribution in **Table 10-1** in **Appendix D**. Compare $Z$ against this critical point. If $Z > z_{cp}$, conclude there is statistically significant evidence at the $\alpha$-level of an increasing trend. If $Z < -z_{cp}$, conclude there is statistically significant evidence of a decreasing trend. If neither exists, conclude that the sample evidence is insufficient to identify a trend.

►EXAMPLE 17-6

Test for a significant upward trend using the Mann-Kendall procedure in the following set of sulfate measurements (ppm) collected over several years.

| Sample No. | Sampling Date (yr.mon) | Sulfate Conc. (ppm) | Sample No. | Sampling Date (yr.mon) | Sulfate Conc. (ppm) |
|---|---|---|---|---|---|
| 1 | 89.6 | 480 | 13 | 93.1 | 590 |
| 2 | 89.8 | 450 | 14 | 93.6 | 550 |
| 3 | 90.1 | 490 | 15 | 94.1 | 600 |
| 4 | 90.3 | 520 | 16 | 94.6 | 700 |
| 5 | 90.6 | 485 | 17 | 95.1 | 570 |
| 6 | 90.8 | 510 | 18 | 95.6 | 610 |
| 7 | 91.1 | 510 | 19 | 95.8 | 650 |
| 8 | 91.3 | 530 | 20 | 96.1 | 620 |
| 9 | 91.6 | 510 | 21 | 96.3 | 830 |
| 10 | 91.8 | 560 | 22 | 96.6 | 720 |
| 11 | 92.1 | 560 | 23 | 96.8 | 590 |
| 12 | 92.6 | 540 | | | |

SOLUTION

Step 1. Construct a time series plot of the sulfate observations to check for a possible trend as in **Figure 17-6**. A clearly rising concentration pattern is seen, although the variability in the measurements appears greater toward the end of the sampling record than at the beginning.

Step 2. Compute the difference between each distinct pair of measurements and determine the sign of the difference, using equation [17.30]. Then sum up the signs with equation [17.31]. Note that to make sure all the distinct pairs have been summed, begin with the first listed observation and compare it to each of values below it. Then take the second listed value and compare it to each of the remaining ones below it, *etc*. The Mann-Kendall statistic becomes:

$$S = \text{sgn}(450-480) + \text{sgn}(490-480) + \ldots + \text{sgn}(590-720) = 196$$

Step 3. Since the sample size $n = 23 > 10$, form the normal approximation to the Mann-Kendall statistic. Because there are some ties in the data, use equation [17.27] to compute the approximate standard deviation. Among the sulfate measurements, there are three groups of ties with 3, 2, and 2 tied values in each set respectively (at values 510, 560, and 590). The adjusted standard deviation is then:

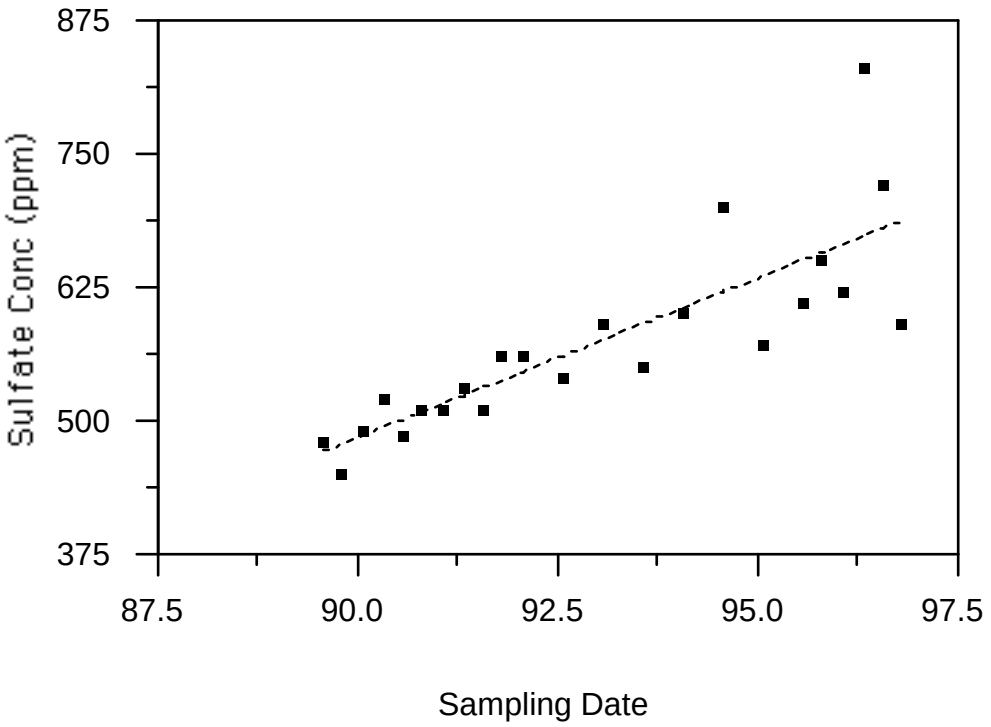
$$SD[S] = \sqrt{\frac{1}{18}\cdot\left[23\cdot(23-1)(2\cdot 23+5) - \left\{3\cdot(3-1)(2\cdot 3+5) + \ldots + 2\cdot(2-1)(2\cdot 2+5)\right\}\right]} = 37.79$$

Finally, using equation [17.29], the normalized Mann-Kendall statistic is:

$$Z = \left(|196| - 1\right) / 37.79 = 5.16$$

Step 4. The $Z$ statistic can be compared to a critical point from the standard normal distribution in **Table 10-1** in **Appendix D**. As large as it is, the test statistic is bigger than the critical point for any usual significance level, suggesting that the trend appears to be real and not just a chance artifact of the sample. ◄

Figure 17-6. Time Series Plot of Sulfate Concentrations (ppm)




## 17.3.3 THEIL-SEN TREND LINE

### BACKGROUND AND PURPOSE

The Mann-Kendall procedure is a non-parametric test for a significant slope in a linear regression of the concentration values plotted against time of sampling. But the Mann-Kendall statistic $S$ does not indicate the *magnitude* of the slope or estimate the trend line itself even when a trend is present. This is slightly different from parametric linear regression, where a test for a significant slope follows naturally from the estimate of the trend line. Even a relatively modest slope can be statistically distinguished from zero with a large enough sample. It is best to first identify whether or not a trend exists, and then determine how steeply the concentration levels are increasing over time for a significant trend. The *Theil-Sen trend line* (Helsel, 2005) is a non-parametric alternative to linear regression which can be used in conjunction with the Mann-Kendall test.

The Theil-Sen method handles non-detects in almost exactly the same manner as the Mann-Kendall test. It assigns each non-detect a common value less than any other detected measurement (*e.g.*,

half the RL). Unlike the Mann-Kendall test, however, the actual concentration values are important in computing the slope estimate in the Theil-Sen procedure. The essential idea is that if a ***simple slope estimate*** is computed for every pair of distinct measurements in the sample (known as the set of ***pairwise slopes***), the average of this series of slope values should approximate the true slope. The Theil-Sen method is non-parametric because instead of taking an ***arithmetic average*** of the pairwise slopes, the ***median*** slope value is determined. By taking the median pairwise slope instead of the mean, extreme pairwise slopes — perhaps due to one or more outliers or other errors — are ignored and have little if any impact on the final slope estimator.

The Theil-Sen trend line is also non-parametric because the median pairwise slope is combined with the median concentration value and the median sample date to construct the final trend line. As a consequence of this construction, the Theil-Sen line estimates the change in ***median*** concentration over time and not the ***mean*** as in linear regression.

## REQUIREMENTS AND ASSUMPTIONS

The Theil-Sen procedure does not require normally-distributed trend residuals as in a linear regression. It is also not critical that the residuals be homoscedastic (***i.e.***, having equal variance over time and with increasing average concentration level). It ***is*** important to have at least 4 and preferably at least 8 or more observation on which to construct the trend. But trend residuals are assumed to be statistically independent. Approximate checks of this assumption can be made using the techniques of **Chapter 14**, once the estimated trend has been removed and the number of non-detect data is limited. Sampling events should also be spaced far enough apart relative to the site-specific groundwater velocity so that an assumption of ***physical*** independence of consecutive sample volumes is reasonable.

A more difficult problem is encountered when a large fraction of the data is non-detect. As long as less than half the measurements are non-detects occurring in the lower part of the observed concentration range, the median concentration value will be quantified and the median pairwise slope will generally be associated with a pair of detects. Larger proportions of non-detect data make computation of the Theil-Sen trend line more difficult and uncertain. The reason is that each time a non-detect is paired with a quantified measurement, the pairwise slope is known only within a range of values. One end of the range results from supposing the true non-detect concentration is equal to zero; the other when the non-detect concentration is equal to the RL.

## PROCEDURE

Step 1. Order the data set by sampling event or time of collection, $x_1$, $x_2$, to $x_n$. Then consider all possible distinct pairs of measurements, ($x_i$, $x_j$) for $j > i$. For each pair, compute the simple pairwise slope estimate:

$$m_{ij} = \left(x_j - x_i\right) \Big/ \left(j - i\right) \qquad [17.32]$$

With a sample size of $n$, there should be a total of $N = n(n–1)/2$ such pairwise estimates $m_{ij}$. If a given observation is a non-detect, use half the RL as its estimated concentration.

Step 2. Order the $N$ pairwise slope estimates ($m_{ij}$) from least to greatest and rename them as $m_{(1)}$, $m_{(2)}$, …, $m_{(N)}$. Then determine the Theil-Sen estimate of slope ($Q$) as the median value of this list.

Finding this value will depend on whether $N$ is even or odd, but the following equation can be used:

$$Q = \begin{cases} m_{(\lceil N+1 \rceil/2)} & \text{if } N \text{ is odd} \\ \left( m_{(N/2)} + m_{(\lceil N+2 \rceil/2)} \right) \Big/ 2 & \text{if } N \text{ is even} \end{cases} \quad [17.33]$$

Step 3. Order the sample by concentration magnitude from least to greatest, $x_{(1)}$, $x_{(2)}$, to $x_{(n)}$. Determine the median concentration with the formula:

$$\tilde{x} = \begin{cases} x_{(n+1)/2} & \textit{if n is odd} \\ \left( x_{n/2} + x_{(n+2)/2} \right) \Big/ 2 & \textit{if n is even} \end{cases} \quad [17.34]$$

Again replace each non-detect by half its RL during this calculation. Also find the median sampling date ($\tilde{t}$) using the ordered times $t_1$, $t_2$, to $t_n$ by a similar computation.

Step 4. Compute the Theil-Sen trend line with the equation:

$$x = \tilde{x} + Q \cdot \left( t - \tilde{t} \right) = \left( \tilde{x} - Q \cdot \tilde{t} \right) + Q \cdot t \quad [17.35]$$

Using equation [17.35], an estimate can be made at any time ($t$) of the expected median concentration ($x$).

►EXAMPLE 17-7

Use the following sodium measurements to compute a Theil-Sen trend line. Note that the sample dates are recorded as the year of collection (2-digit format) plus a fractional part indicating when during the year the sample was collected. This allows an annual slope estimate, since 1 unit = 1 year.

| Sample Date (yr) | Sodium Conc. (ppm) |
|---|---|
| 89.6 | 56 |
| 90.1 | 53 |
| 90.8 | 51 |
| 91.1 | 55 |
| 92.1 | 52 |
| 93.1 | 60 |
| 94.1 | 62 |
| 95.6 | 59 |
| 96.1 | 61 |
| 96.3 | 63 |

SOLUTION

Step 1. Compute the pairwise slopes for each distinct pair of measurements using equation [17.32]. With $n$ = 10 observations, there will be a total of 10(9)/2 = 45 such pairs. The first few are listed below:

$$m_{12} = (53-56)/(90.1-89.6) = -6$$
$$m_{13} = (51-56)/(90.8-89.6) = -4.17$$
$$m_{14} = (55-56)/(91.1-89.6) = -.667$$

Step 2. Since the total number of distinct pairs is odd, sort the list of pairwise slopes as in the table below and let Sen's estimated slope equal the middle or 23rd largest value in this list. This gives an estimate of $Q$ = 1.33 ppm increase *per year*, an estimate in line with the time series plot of **Figure 17-7**.

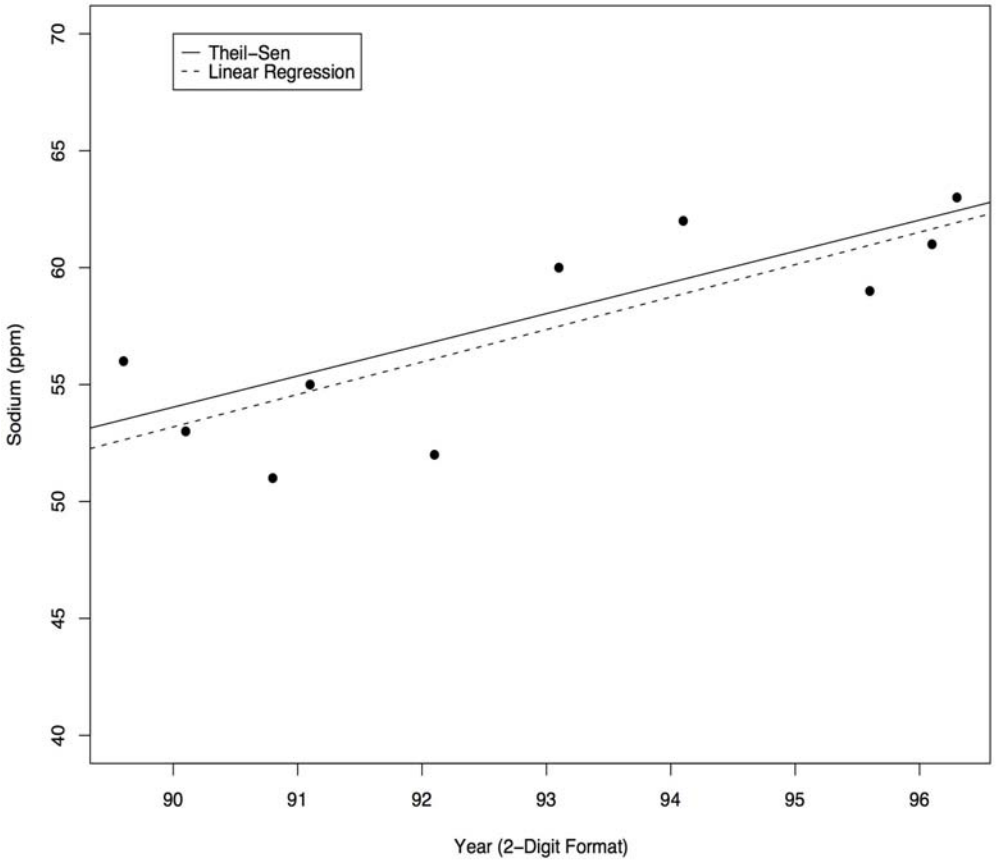
Step 3. Compute the median concentration value $\tilde{x} = 57.5$ and the median sample date $\tilde{t} = 92.6$ from the table above. Then calculate the Theil-Sen trend line using the slope estimate from Step 2:

$$x = 57.5 + 1.333(t - 92.6) = -65.97 + 1.333t$$

This trend line can be used to estimate the predicted median concentration ($x$) at any desired time in years ($t$). For example, at the beginning of 1998 ($t$ = 98), the trend line would predict a median sodium concentration of approximately $x$ = 64.7 ppm. ◄

| Rank | Pairwise Slope | Rank | Pairwise Slope |
|---|---|---|---|
| 1 | -6 | 24 | 1.538 |
| 2 | -4.167 | 25 | 1.613 |
| 3 | -3 | 26 | 1.667 |
| 4 | -2.857 | 27 | 1.887 |
| 5 | -2 | 28 | 2 |
| 6 | -1.6 | 29 | 2 |
| 7 | -0.667 | 30 | 2 |
| 8 | -0.5 | 31 | 2.182 |
| 9 | -0.5 | 32 | 2.25 |
| 10 | -0.4 | 33 | 2.25 |
| 11 | 0.333 | 34 | 2.333 |
| 12 | 0.455 | 35 | 2.333 |
| 13 | 0.5 | 36 | 2.5 |
| 14 | 0.769 | 37 | 2.619 |
| 15 | 0.769 | 38 | 3.333 |
| 16 | 0.889 | 39 | 3.913 |
| 17 | 0.938 | 40 | 4 |
| 18 | 1.045 | 41 | 5 |
| 19 | 1.091 | 42 | 5.714 |
| 20 | 1.143 | 43 | 8 |
| 21 | 1.2 | 44 | 10 |
| 22 | 1.333 | 45 | 13.333 |
| 23 | 1.333 | | |

Figure 17-7. Time Series Plot of Sodium Concentrations (ppm)



# CHAPTER 18. PREDICTION LIMIT PRIMER



This chapter introduces the concept of statistical intervals and focuses on several types of prediction limits useful for detection monitoring. The requirements and common assumptions of such limits are explained, as well as specific descriptions of:

- Prediction limits for *m* future values (**Section 18.2.1**)
- Prediction limits for future means (**Section 18.2.2**)
- Non-parametric prediction limits for *m* future values (**Section 18.3.1**)
- Non-parametric prediction limits for a future median (**Section 18.3.2**)

## 18.1 INTRODUCTION TO PREDICTION LIMITS

First discussed in **Chapter 6,** *prediction limits* belong to a class of methods known as *statistical intervals*. Statistical intervals represent concentration or measurement ranges computed from a sample that are designed to estimate one or more characteristics of the parent population. In groundwater monitoring, statistical intervals offer a convenient and statistically valid way to test for significant differences between background versus compliance point groundwater measurements.

The statistical interval accounts for variability inherent not only in future measurements, but also additional uncertainty in the prediction limit itself. The latter is derived from a relatively small background sample with an associated level of variability in estimating the true characteristics of the underlying groundwater population.

Prediction limits are generally easy to construct and have a straightforward interpretation. Background data are used to construct a concentration limit *PL*, which is then compared to one or more observations from a compliance point population. The acceptable range of concentrations includes all values no greater than the prediction limit. The appropriate *prediction interval* will generally have the form [0, *PL*], with the upper limit *PL* as the comparison of importance. Unless pH or a similar parameter is being monitored, a one-sided upper prediction limit is used in detection monitoring.

A significant advantage to prediction limits is their *flexibility*, which can accommodate a wide variety of groundwater monitoring networks. Prediction limits can be constructed so that as few as *one*

new measurement per compliance well may suffice for a test. Prediction limits may be based on a comparison of means, medians, or individual compliance point measurements, depending on the characteristics of the monitoring network and the constituents being tested.

Prediction limits can also be designed to accommodate a wide range of *multiple statistical comparisons* or *tests*. Each periodic statistical evaluation (*e.g.*, semi-annually) under RCRA and other regulations involves separate tests at all compliance well locations for each monitoring constituent. Often, the number of separate statistical tests can be quite sizeable. Prediction limits can be constructed to precisely account for the number of tests to be conducted, so as to limit the *site-wide false positive rate* [SWFPR] and ensure an adequate level of *statistical power* (see discussion in **Chapter 6**).

This and the following chapter present basic concepts and procedures for using prediction limits as detection monitoring tests. The intent is to provide a relatively simple framework for using prediction limits in RCRA or CERCLA groundwater monitoring. **Chapter 18** describes the construction of prediction limits for tests involving a single constituent at one well. It describes the basic mechanics of each type of prediction limit and how they differ from one another.

**Chapter 19** expands this discussion to cover multiple *simultaneous* prediction limit tests (*i.e.*, all occurring during a single statistical evaluation or during a single year of monitoring). Cumulative SWFPRs and statistical power are considered, including how these criteria impact the expected performance of a given prediction limit strategy. Examples are provided to illustrate these procedures, as well as explanations of associated tables and software.

Specific strategies in **Chapter 19** apply the concept of *retesting*. Generally speaking, ***almost any prediction limit procedure in detection monitoring should be combined with an appropriate retesting strategy***. The reason is that when testing a large number of compliance point samples, it is almost guaranteed that one or more measurements will exceed an upper prediction limit. *Resampling* of those wells where an exceedance has occurred can either verify the initial evidence of a release or disconfirm it, while avoiding unnecessary false positives.

**Chapter 6** introduced a number of key terms used in the Unified Guidance, especially for prediction limit and control chart tests. The guidance applies the term *comparison* to individual future measurements or sample statistics evaluated against a prediction limit (or *control chart limit*), and the term *test* to represent a series of future data comparisons that ultimately result in a statistical decision. A 1-of-*m* retesting procedure (described below), for instance, might involve comparison of up to *m* distinct sample measurements against the prediction limit. Each of these individual samples involves a *comparison*, but only after all the necessary individual comparisons have been made is the *test* complete. This distinction becomes particularly important when properly determining SWFPRs, a subject discussed both in **Chapter 6** and **Chapter 19**.

One or more *future* observations are collected for purposes of testing compliance well data, as distinct from the *background* sample from which the prediction limit is constructed. Background data can be obtained from upgradient wells or in combination with historical, uncontaminated compliance well data. In intrawell testing, data from an individual compliance well constitute both the background and future samples. The two data sets need to be distinct and may not overlap, even if the historical

background data is periodically updated with previously evaluated future samples. The key idea is that at any given point in time, background and future data sets are clearly distinguished.

Formally, prediction limits are constructed to contain one or more future observations or sample statistics generated from the background population with a specified probability equal to (1–α). The probability (1–α) is known as the confidence level of the limit. It represents the chance — over repeated applications of the limit to many similar data sets — that the prediction limit will contain future observations or statistics drawn from its background population.

A sample of *n* background measurements is used to construct the prediction limit. Under the null hypothesis that the compliance point population is identical to background, a set of *m* independent compliance point observations or a statistic like the mean based on those observations (*i.e.*, the future data) is then compared against the prediction limit. For the prediction limit to serve as a valid statistical test, the future observations are initially presumed to follow the same distribution as background.

Only background values are used to construct the prediction limit. But the probability that the limit contains all *m* future observations or sample statistics derived from those future data does not depend solely on the observed background. It is also based on the number of future measurements or sample statistics used in the comparison and *how* the individual comparisons are conducted. To underscore this point, consider the general equation for a prediction limit based on normal or transformably normal populations, given by

$$PL = \bar{x} + \kappa s \qquad [18.1]$$

where $\bar{x}$ is the sample mean in background, $s$ is the background standard deviation, and $\kappa$ is a multiplier depending on the type of prediction limit under construction. The simplest type of prediction limit test compares a specific number of individual future observations to the limit (*PL*). For example, do all three compliance measurements collected during a 6-month period fall within the prediction interval? The multiplier $\kappa$ and hence the prediction limit itself, changes depending on whether one, two or three compliance observations will be compared against *PL*. More generally, the κ-multiplier is selected to account not only for the number of future comparisons, but also for the *rules of the comparison strategy* and the number of simultaneous tests to be conducted (*e.g.*, the number of monitoring constituents times the number of compliance wells).

In the simplest case of a successive comparison of *m* individual future measurements against *PL*, the test is labeled as an *m*-of-*m* prediction limit. All *m* of the future observations need to fall within the prediction interval for the test to 'pass' — that is, be no greater than *PL*. If any one or more of the future values exceed the PL, the test fails and the well is deemed to have a *statistically significant increase* [SSI] or constitute an exceedance.

The $\kappa$-multiplier appropriate for an *m*-of-*m* prediction limit test is different from the multipliers that would be computed for other kinds of comparison rules. Another simple type is a comparison of a single future *mean of order p*. Here, *p* future measurements are collected and *averaged* before comparing against *PL*. If the order-*p* mean is no greater than *PL*, the test passes; otherwise, it fails. A test following this rule is labeled a *1-of-1 prediction limit on a future mean*. The important thing to remember is that the $\kappa$-multiplier and thus the prediction limit will differ depending on whether or not the *p* future values are first averaged or simply compared against *PL* one-by-one. The choice to use one rule versus the other

impacts the magnitude of the prediction limit and ultimately its expected statistical power and false positive rate.

Other comparison rules of substantial benefit in groundwater monitoring are 1-of-*m* prediction limit on future observations or a statistic like the mean or median. This test requires at least one of *m* successive observations or statistic to fall within the prediction interval in order to pass. Operationally this means that if an initial compliance well measurement is no greater than *PL*, the test is complete and no further sampling need be done. If the initial value exceeds *PL*, one or more of (*m*–1) *resamples* need to be obtained. Since these additional measurements are collected sequentially over sufficiently long time periods to maintain approximate statistical independence (**Chapter 3**), the first resample to fall within the prediction interval also ends the test as 'inbounds' or passing, frequently obviating the need to gather all *m* measurements.

Another comparison rule of some use is known as the California strategy, first developed for the State of California RCRA program. The California strategy can be construed as a *conditional* rule: if an initial future observation is no greater than *PL*, further comparisons are not needed and the test passes. However, if the initial observation exceeds the *PL*, 2-of-2 or 3-of-3 resamples *all need to not exceed the PL* in order for the well to remain in compliance. A slight modification to this rule termed the *modified California* approach has better statistical power and false positive rate characteristics than the original California strategies, and is therefore included as a potential prediction limit test.

### 18.1.1 BASIC REQUIREMENTS FOR PREDICTION LIMITS

All prediction limits share certain basic assumptions when applied as tests of groundwater. Further, *parametric* prediction limits as presented in the Unified Guidance require the sample data to be either normally-distributed or normalized via a transformation. The key points can be summarized as follows:

1. background and future sample measurements need to be identically and independently distributed (the *i.i.d.* presumption; see **Chapter 3**);
2. sample data do not exhibit temporal non-stationarity in the form of trends, autocorrelation, or other seasonal or cyclic variation;
3. for interwell tests (*e.g.*, upgradient-to-downgradient comparisons), sample data do not exhibit non-stationary distributions in the form of significant natural spatial variability;
4. background data do not include statistical outliers (a form of non-identical distributions);
5. for parametric prediction limits, background data are normal or can be normalized using a transformation; and
6. a minimum of 8 background measurements is available; more for non-parametric limits or when accounting for multiple, simultaneous prediction limit tests.

The first assumption implies that background data are randomly drawn from a single common parent population, especially if aggregated from more than one source well. As discussed in **Chapter 5**, analysis of variance [ANOVA] can be used to determine the appropriateness of pooling data from

different background wells. There is also a presumption that the compliance point measurements follow the same distribution as background in the absence of a release.

The second assumption is corollary to the first, and requires that the background data are *stationary over time* (**Chapter 3**). This can be evaluated with one or more techniques described in **Chapter 14** on temporal variability. These account for trends, autocorrelation, or other variation, perhaps by utilizing *data residuals* instead of the raw measurements. If the background residuals meet the basic points above, they can be used to construct an adjusted prediction limit. Residuals of the future observations would also need to be computed and compared against the adjusted prediction limit to ensure a valid and consistent test.

The second assumption also requires that there be only a single source of variation in the data, when using the usual sample standard deviation (*s*) to compute the prediction limit. If there are other sources of variation such as seasonal patterns or temporal variation in lab analytical performance, these should be included in the estimate of variability. Otherwise *s* is likely to be *biased*. One method to accomplish this is by use of an appropriate ANOVA model to include temporal factors affecting the variability (**Chapter 14**). Determination of the components of variance in more complicated models is beyond the scope of this guidance and may require consultation with a professional statistician.

The third assumption requires that background and compliance point populations be identical in distribution, absent a release, for interwell tests. Spatial variation violates this assumption since the well population means ($\mu$) will be different, making it impossible to know whether an apparent upgradient-to-downgradient difference is attributable to a release or simply variations in natural groundwater concentration levels. The assumption also requires that each population share a common variance ($\sigma^2$). Tests of equal variance (*i.e.*, homoscedasticity) when using prediction limits may be possible either by examining groups of historical background and compliance point data or by performing periodic tests when enough compliance point measurements have been accumulated to make a diagnostic test possible.

The fourth assumption implies that background data should be screened for outliers using the techniques in **Chapter 12**. Statistical outliers can potentially inflate a prediction limit and severely limit its statistical power and accuracy by over-inflating both the sample background mean ($\bar{x}$) and especially the background standard deviation (*s*). The Unified Guidance discourages automated removal of outliers from background samples, but all possible outliers should be examined to determine whether a cause can be identified (see discussion in **Chapter 6**). In some cases, an apparent outlier may represent a valid portion of the underlying background population that has not yet been sampled or observed. It also could represent evidence that conditions in background have changed or are changing.

The fifth assumption of normality for parametric prediction limits can be evaluated using the diagnostic techniques described in **Part II** of the guidance. If skewed background data can be normalized via a transformation (*e.g.*, the natural logarithm), the prediction limit should be constructed on the transformed background values. The resulting limit should either be: 1) back-transformed to the concentration domain (*e.g.*, by exponentiation) when comparing future individual compliance observations; or 2) left in the transformed scale when compared to future mean compliance data also based on *the same transformation*. In the latter case, use of a logarithmic transformation results in evaluating population medians or geometric means and *not* the arithmetic means.

When normality cannot be justified, a non-parametric prediction limit should be considered instead. A non-parametric limit assumes only that all the data come from the same, usually unknown, continuous population. Non-parametric prediction limits generally require a much larger number of background observations in order to provide the same level of confidence (1–α) as a comparable parametric limit. Consequently, the Unified Guidance recommends that a parametric model be fit to the data if at all possible.

The last assumption concerns sufficient background sample sizes. A prediction interval can be computed with as few as three observations from background. However, this can result in an unacceptably large upper prediction limit and a test with very limited statistical power. A sample size of eight or more is generally needed to derive an adequate parametric prediction limit, especially if a retesting strategy is not employed. The exact requirements depend on the number of simultaneous tests (*i.e.*, number of wells times number of constituents per well) to be made against the prediction limit and the type of retesting strategy adopted (see **Chapter 19** for more discussion of retesting strategies).

If a minimum schedule of quarterly sampling is being followed and there is only one background well, at least two years of data will be needed before constructing the prediction limit.[1] If data from multiple background wells screened in comparable hydrologic conditions can reasonably be combined (see **Chapter 5**), pooling background data to increase background sample sizes is encouraged.

### 18.1.2 PREDICTION LIMITS WITH CENSORED DATA

When a sample contains a substantial fraction of non-detects or left-censored measurements, it may be impossible to even approximately normalize the data A sample data set may originate from a normal or transformable-to-normal population, but the uncertainty surrounding both the censored values and the consequent shape of the lower tail of the distribution prevents a clear identification. If the apparent underlying distribution is not normal or transformable to normality, a non-parametric prediction limit (**Section 18.3**) should be used.

Given that non-parametric prediction limits typically have much steeper background data requirements than their parametric counterparts, one remedy is to attempt a fit to normality by using censored probability plots (**Chapter 15**) in conjunction with either the *Kaplan-Meier* or *robust regression on order statistics* [ROS] techniques (**Chapter 15**) for left-censored data. Censored observations prevent a full and complete ordering of the sample, making it difficult to assess normality with standard probability plots (**Chapter 9**). Censored probability plots, on the other hand, only graph the detected values, but do so based on a *partial ordering and ranking* of the sample. Data that appear distinctly non-normal on a standard probability plot (where non-detects are perhaps replaced by half their reporting limits [RLs] to allow plotting) can sometimes appear reasonably normal on a censored probability plot. Transformations can also be applied and the censored probability plot reconstructed to see if the data can be normalized in that fashion.

[1] The Unified Guidance does not recommend that only one background well be used in any kind of interwell or upgradient-to-downgradient comparison. Multiple background wells are always preferred so that tests for spatial variability may be made and the exact nature of background better understood.

If the censored probability plot is close to linear and the sample approximately normalized, an estimated mean and standard deviation should be computed. These estimates will not be the same if each non-detect were replaced by half its RL, and the sample mean calculated from the resulting imputed sample. To properly account for the censoring, the estimated mean (denoted as $\hat{\mu}$) and the estimated standard deviation ($\hat{\sigma}$) needs to be derived as parameters from the normal distribution providing the closest fit to a partial ordering of the sample (as on a censored probability plot). The Unified Guidance describes two slightly different techniques for accomplishing this task.

Once $\hat{\mu}$ and $\hat{\sigma}$ estimates have been computed, an adjusted parametric prediction limit is constructed by substituting $\hat{\mu}$ for $\bar{x}$ and $\hat{\sigma}$ for s in the equations of **Section 18.2** or **Chapter 19**. For example, the adjusted equation for a general parametric prediction limit would become:

$$PL = \hat{\mu} + \kappa \cdot \hat{\sigma} \qquad [18.2]$$

Another potential difference between the adjusted prediction limit in equation [18.2] and the unadjusted prediction limit in equation [18.1] is the number of *degrees of freedom* [*df*] used in selecting the $\kappa$-multiplier. Absent any censored measurements, a background sample of size *n* would normally have (*n*–1) *df*. With censoring, there is greater statistical uncertainty surrounding each non-detect than surrounding the detected values. Because of this, the actual degrees of freedom is somewhere between *d* (the number of detects) and (*n*–1) (the total sample minus one). Unfortunately, there is no straightforward, general method to determine the true *df*. To be conservative, the *df* should be set equal to *d*, since the value of each detect is known with reasonable certainty. Setting a lower *df* tends to raise the $\kappa$-multiplier and thus the prediction limit over what would be selected with an uncensored sample of the same size. This is consistent with the greater uncertainty associated with non-detect measurements. However, it is at best an approximate remedy. Further consultation with a professional statistician may be warranted to arrive at a better choice of the degrees of freedom.

# 18.2 PARAMETRIC PREDICTION LIMITS

## 18.2.1 PREDICTION LIMIT FOR M FUTURE VALUES

### BACKGROUND AND PURPOSE

A prediction limit test for *m* future values is constructed so that *m* compliance point observations are evaluated by determining whether or not they fall within a prediction interval derived from background. As mentioned in **Chapter 2**, some State programs may require up to 4 successive sampling events per evaluation period for testing, which can be addressed by the prediction limit approach described below.

If the distributions of background and compliance point data are identical as assumed under the null hypothesis $H_0$, *all m of the compliance point observations* should be no greater than the upper prediction limit [*PL*]. If any of the future observations exceeds *PL*, there is statistical evidence that the

compliance data do not come from the same distribution as background, but instead are elevated above background.[2]

With intrawell comparisons, a prediction limit can be computed on historical data or intrawell background to contain a specified number (*m*) of future (*i.e.*, more recent) observations from the same well. If any of the future values exceeds the upper prediction limit, there is evidence of recent contamination at the well.

## REQUIREMENTS AND ASSUMPTIONS

As noted in **Section 18.1**, the prediction limit test on *m* future values is designated as an *m*-of-*m* test. *Each* of the *m* individual future observations need to be compared to the prediction limit [*PL*]. All should be no greater than *PL* for the test to pass. The number of future observations to be collected (*m*) need to be specified *in advance* in order to correctly compute the κ-multiplier from equation [18.1]. Consequently, if compliance data are collected on a regular schedule, the prediction interval can be constructed to cover a specified *time period* of future sampling. Usually this period will coincide with the time between statistical evaluations specified in the site permit (*e.g.*, on a semi-annual or annual basis). Keep in mind also that *m* denotes the number of consecutive sampling events being compared to the prediction limit at a given well for a given constituent.

As discussed in more detail in **Chapter 6**, a new prediction limit should be constructed prior to each statistical evaluation for *interwell* tests, when additional background data have been collected along with the new compliance point measurements. Unless there is evidence of characteristic changes within background groundwater quality (*e.g.*, as demonstrated by observable trends in background), background data should be amassed or accumulated over time. Earlier background measurements need not be discarded, both to maintain an adequate background sample size and also because a larger span of sampling results will provide a better statistical description of the underlying background population. The revised prediction limit will then reflect a larger background sample size, *n*, but possibly the same number, *m*, of future values to be predicted at the next statistical evaluation.

For *intrawell* tests, the prediction limits should be revised only after intrawell background has been updated (**Chapter 5**). Such updating may not coincide with the regular schedule of statistical evaluations if done, for instance, every two years or so. In that case, the same intrawell prediction limit might be used for multiple evaluations before being revised.

## PROCEDURE

Step 1. Calculate the sample mean $\bar{x}$, and standard deviation *s*, from the set of *n* background measurements.

Step 2. Specify the number of individual future observations (*m*) from the compliance well to be included in the prediction interval for an *m*-of-*m* test. For an upper prediction limit with an overall $(1-\alpha)$ confidence test level for the *m* comparisons, use the equation:

[2] In the context of the Unified Guidance, *m* represents the number of consecutive samples being compared in the prediction limit test for a given well and constituent.

$$PL = \bar{x} + t_{1-\alpha/m,n-1} s \sqrt{1+\frac{1}{n}} \quad [18.3]$$

It is assumed that exactly $m$ consecutive sample values from the compliance point will be compared against the upper $PL$. Note that the quantile from a Student's $t$-distribution used in equation [18.3] has two parameters: the degrees of freedom ($n$–1) and a joint comparison confidence level $\left(1-\alpha/m\right)$. Most Student's $t$-quantiles can be found directly or approximated through interpolation by looking in **Table 16-1** of **Appendix D**.

Note: equation [18.3] assumes the prediction limit is applied to only one constituent at a single well. If multiple tests need to be performed (e.g., on multiple wells and/or multiple constituents), the prediction limit takes the form:

$$PL = \bar{x} + \kappa s \quad [18.4]$$

where the κ-multiplier is determined using one of the strategies described in **Chapter 19**.

If a log transformation is applied to the data to bring about approximate normality, the upper $PL$ should be constructed using the log-mean ($\bar{y}$) and log-standard deviation ($s_y$), using the equation:

$$PL = \exp\left( \bar{y} + t_{1-\alpha/m,n-1} s_y \sqrt{1+\frac{1}{n}} \right) \quad [18.5]$$

If multiple tests must be conducted and a log transformation has been applied to the data, the upper $PL$ will have the form:

$$PL = \exp\left(\bar{y} + \kappa s_y\right) \quad [18.6]$$

Note: other transformations besides the natural logarithm are handled in a similar manner; compute the prediction limit on the transformed data, then back-transform the limit to the original concentration scale prior to comparison with any future observations.

Step 3. Once the prediction limit ($PL$) has been calculated, compare each of $m$ compliance point future values against $PL$. If all of these measurements are no greater than $PL$, the test passes and the well is deemed to be in compliance. If, however, any compliance point concentration exceeds $PL$, there is statistically significant evidence of an increase over background.

### ►EXAMPLE 18-1

The data in the table below represent quarterly arsenic concentrations measured in a single well at a solid waste landfill. Calculate an intrawell upper prediction limit for 4 future samples with 95% confidence and determine whether there is evidence at the annual statistical evaluation of a possible release during Year 4 of monitoring.

| **Intrawell Background** | | **Compliance Data** | |
|---|---|---|---|
| Sampling Period | Arsenic (ppb) | Sampling Period | Arsenic (ppb) |
| Year 1 | 12.6 | Year 4 | 48.0 |
| | 30.8 | | 30.3 |
| | 52.0 | | 42.5 |
| | 28.1 | | 15.0 |
| Year 2 | 33.3 | | |
| | 44.0 | | |
| | 3.0 | | |
| | 12.8 | | |
| Year 3 | 58.1 | | |
| | 12.6 | | |
| | 17.6 | | |
| | 25.3 | | |
| | $n$ = 12 | | |
| | Mean = 27.52 | | |
| | SD = 17.10 | | |

SOLUTION

Step 1. First check the sample data for the key points identified in **Section 18.1.1**. As an example, a Shapiro-Wilk test on the background data gives a test statistic of $SW$ = 0.947. The critical point at the $\alpha$ = .05 level for the Shapiro-Wilk test on $n$ = 12 observations is 0.859. Since the test statistic exceeds the critical point, there is insufficient evidence to reject an assumption of normality.

Step 2. Compute the prediction interval using the raw background data. The sample mean and standard deviation of the 12 background samples are 27.52 ppb and 17.10 ppb, respectively.

Step 3. A single future year of compliance data then is compared to the prediction limit, leading to a test of $m$ = 4 individual measurements. Setting the overall confidence level to $(1-\alpha)$ = 95%, the probability used to determine an appropriate Student's $t$-quantile needs to be set to $(1-\alpha/m) = 1-.05/4 = .9875$. The $t$-distribution with probability .9875 and (n–1) = 11 degrees of freedom in **Table 16-1** of **Appendix D** results in a $t$-quantile of 2.593. Using equation [18.3], the upper prediction limit can be computed as:

$$PL = 27.52 + t_{.9875,11}(17.10)\sqrt{1+\frac{1}{12}} = 27.52 + 2.593(17.10)\sqrt{1.0833} = 73.67 \text{ ppb}$$

Step 4. Compare the upper $PL$ to each compliance measurement in Year 4. None of the four observations exceeds 73.67 ppb. Consequently, there is no statistically significant evidence of arsenic contamination during that year. ◄

## 18.2.2 PREDICTION LIMIT FOR A FUTURE MEAN

### BACKGROUND AND PURPOSE

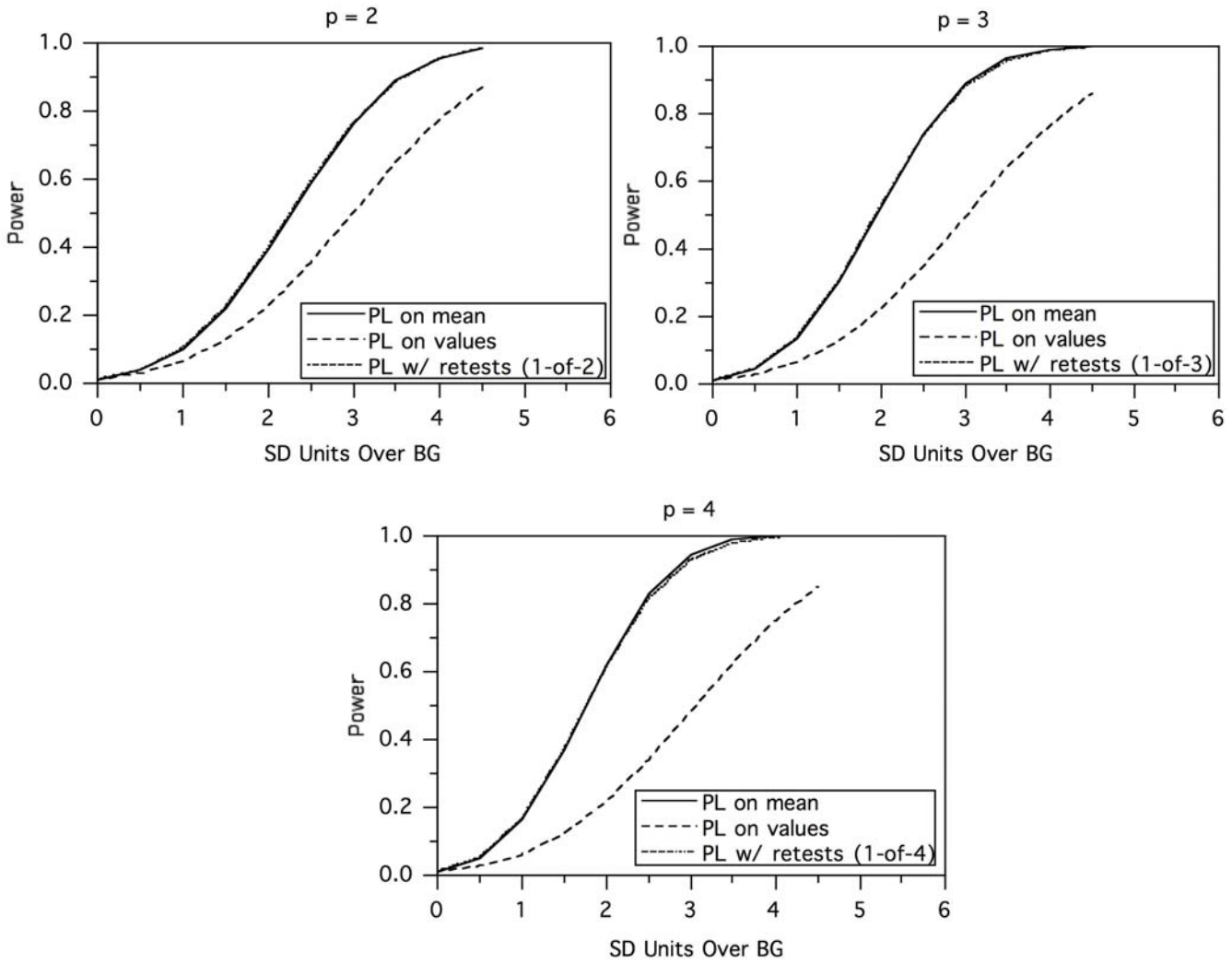
Although prediction limits are often constructed as bounds on extreme individual measurements, they can also be formulated to predict an acceptable range of concentrations for the *mean* of *p* future values. The comparison rule for the test is then different: instead of requiring all of a set of m individual values to fall within the prediction interval for the test to pass, only the *average* of the (*p*) future values should not exceed the prediction limit.

In this setting, the prediction limit for a future mean is more nearly akin to a *t*-test or parametric ANOVA, since the mean of the compliance point well is compared to a limit based on the background mean. The principal differences in using a prediction limit as opposed to those tests are: first that the variability of the compliance point population is *assumed* to be identical to that in background. With a *t*-test or ANOVA, each distinct well group contributes to the overall estimate of variability, not merely the background values. Secondly, t-tests and especially ANOVA are typically utilized as interwell tests, whereas prediction limits for a future mean can be constructed for either interwell or intrawell testing.

The hypothesis being tested when using a prediction limit for a future mean in detection monitoring is exactly the same as that posited for a prediction limit for *m* future values, namely, $H_0$: background population is identical to compliance population (implying $\mu_C \leq \mu_{BG}$) *vs*. $H_A$: compliance mean is greater than background mean (*i.e.*, $\mu_{DG} > \mu_{BG}$). However, the statistical properties of the two prediction interval formulations are somewhat different.

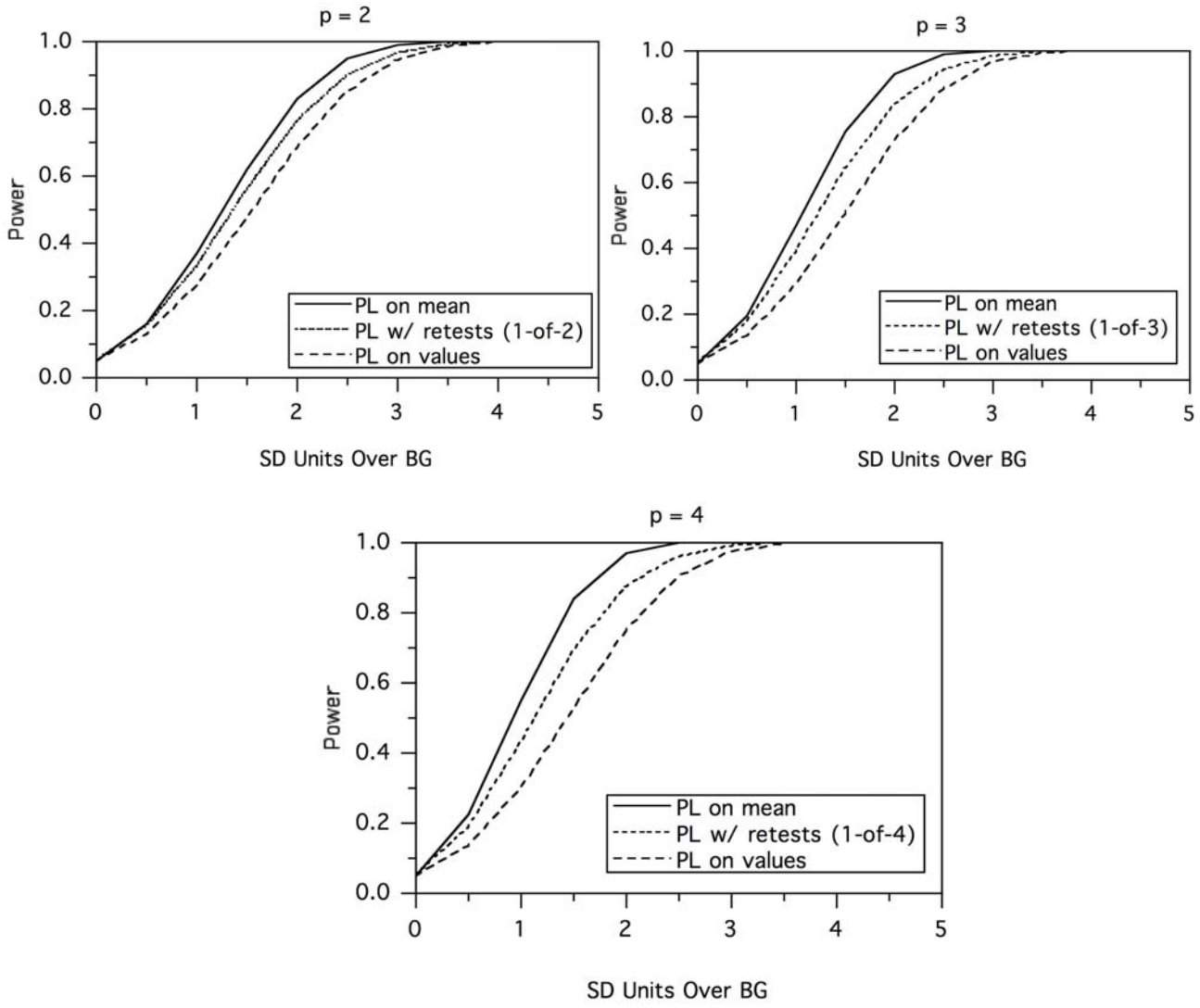
For the same background sample size (*n*), false positive rate ($\alpha$), and number of future samples where $p = m$, the power of the prediction limit for a future mean of order *p* with normally-distributed data is generally higher than for a prediction limit of the next *m* individual future observations. *This suggests that* ***when feasible and appropriately implemented,*** *a prediction limit strategy based on future means may be more environmentally protective than a strategy based on predicting individual future measurements*. A few examples of the power differences are presented in **Figures 18-1** and **18-2**.

Figure 18-1. Comparison of Prediction limits (BG = 8, $\alpha$ = .01, 1 test)




Even when a retesting strategy is employed, such as the 1-of-*m* schemes for prediction limits on individual values described in **Chapter 19**, the statistical power at best *matches* that of a prediction limit on a single future mean with no retesting, when the same numbers of background and compliance point measurements are used. As **Figure 18-2** illustrates, for some cases the 1-of-*m* power is comparatively lower. Under background conditions, 1-of-*m* strategies provide an earlier indication of *uncontaminated* groundwater, since a single observation can indicate uncontaminated groundwater. By contrast, all $p = m$ individual samples need to be collected to form a mean of order $p = m$ when using a prediction limit test for a single future mean. With a groundwater release, no such potential time savings exists. In that case, *all p or m* samples need to be collected with either type of prediction limit.

Figure 18-2. Comparison of Prediction limits (BG = 20, $\alpha$ = .05, 1 test)




## REQUIREMENTS AND ASSUMPTIONS

Although a prediction limit for a future mean is generally preferable in terms of statistical power for identifying potential contamination, it is not always practical to implement. To accommodate the large number of statistical tests that all but the smallest sites must contend with, the Unified Guidance recommends that almost any prediction limit be implemented in conjunction with a retesting strategy (**Chapter 19**). Otherwise, the prediction limit formulations provided in this chapter will likely fall short of providing an adequate balance between false negative and positive decision errors. Retesting with a prediction limit for a future mean will necessitate the collection of *p* additional measurements to form the resampled mean, whenever the initial future mean exceeds the prediction limit. Since all prediction limit tests assume that both the background and compliance data are statistically independent, there needs to generally be enough temporal spacing between sampling events to avoid introducing significant autocorrelation in the series of compliance point values.

If semi-annual evaluation of groundwater quality is required, and depending on data characteristics (see **Chapter 14** discussions on temporal variability), there may not be sufficient time for collecting at least 4 independent groundwater measurements from a given well over a six-month period. This would be the minimum needed to form an initial mean and potentially a resample mean of order 2. To avoid this dilemma, the guidance discusses an alternate approach in **Chapter 19** for using 1-of-1 prediction limit tests on means.

Like the parametric prediction limit for *m* future values, the prediction limit on a future mean assumes that the background data used to construct the limit are either normally-distributed or can be normalized. If a transformation is used (*e.g.*, the natural logarithm) and the limit built on the transformed values, the prediction limit should *not* be back-transformed before comparing to the compliance point data. Rather, because of transformation bias in the mean, the compliance point data should also be transformed, and the future mean computed from the *transformed* compliance measurements. Then the *mean of the transformed values* (*e.g.*, log-mean) should be compared to the prediction limit in the transformed domain. As previously mentioned, the prediction limit in the logarithmic domain is not a test of the arithmetic mean, but rather of the *geometric mean* or *median* (also see **Chapter 16**). In most situations, a decision that the lognormal median at the compliance point exceeds background will also imply that the lognormal arithmetic mean exceeds background.

## PROCEDURE

Step 1. Calculate the sample mean, $\bar{x}$, and the standard deviation, *s*, from the set of *n* background measurements.

Step 2. Specify the order (*p*) of the mean to be predicted (*i.e.*, the number of individual compliance observations to be averaged). If the background data are approximately normal and an upper prediction limit with confidence level (1−α) is desired, use the equation:

$$PL = \bar{x} + t_{1-\alpha,n-1} s \sqrt{\frac{1}{p} + \frac{1}{n}} \qquad [18.7]$$

where it is assumed that an average of *p* consecutive sample values from the compliance point will be compared against *PL*. Note that the Student's *t*-quantile used in the equation has two parameters: the degrees of freedom (*n*–1) and the cumulative probability (1−α). Most Student's *t*-quantile values can be found directly or approximated through interpolation by using **Table 16-1** in **Appendix D**.

Note also that equation [18.7] assumes that the prediction limit is applied to only one constituent at a single well. If multiple tests are to be conducted and a retesting procedure is employed, the prediction limit will take the form of equation [18.4] where the κ-multiplier is determined using the tables described in **Chapter 19**.

Step 3. If a log transformation is applied to normalize the background sample, the upper *PL* on the log-scale should be constructed using the log-mean ($\bar{y}$) and log-standard deviation ($s_y$), using equation [18.8]:

$$PL = \bar{y} + t_{1-\alpha,n-1} s_y \sqrt{\frac{1}{p} + \frac{1}{n}} \quad [18.8]$$

Note that unlike the lognormal prediction limit for future values, the limit in equation [18.8] is not exponentiated back to the concentration domain. Also, equation [18.8] only applies to a single test (*i.e.*, one constituent at a single well). If multiple tests are to be performed, the prediction limit will have the form:

$$PL = \bar{y} + \kappa s_y \quad [18.9]$$

where the κ-multiplier is again determined from the tables described in **Chapter 19**.

Other transformations are handled similarly: construct the prediction limit on the transformed background, but do *not* back-transform the limit.

Step 4. Once the limit has been computed, compare the compliance point mean against the prediction limit. If the compliance point mean is below the upper *PL*, the test passes. If the mean exceeds the *PL*, there is statistically significant evidence of an increase over background.

## ►EXAMPLE 18-2

The table below contains chrysene concentration data found in water samples obtained from two background wells (Wells 1 and 2) and a compliance well (Well 3). Compute the upper prediction limit for a future mean of order 4 with 99% confidence and determine whether there is evidence of possible chrysene contamination.

| | **Chrysene (ppb)** | | | |
|---|---|---|---|---|
| | Background | | | Compliance |
| Month | Well 1 | Well 2 | Joint | Well 3 |
| 1 | 6.9 | 15.1 | | 68.0 |
| 2 | 27.3 | 7.2 | | 48.9 |
| 3 | 10.8 | 48.4 | | 30.1 |
| 4 | 8.9 | 7.8 | | 38.1 |
| Mean | 13.47 | 19.62 | 16.55 | 46.28 |
| SD | 9.35 | 19.52 | 14.54 | 16.40 |
| Log-mean | 2.451 | 2.656 | 2.533 | 3.789 |
| Log-SD | 0.599 | .881 | .706 | 0.349 |

### SOLUTION

Step 1. Before constructing the prediction limit, check the key assumptions. Assuming there is no substantial natural spatial variability and it is appropriate to combine the background wells into a single data pool, the algorithm for a parametric prediction limit presumes that the background data jointly originate from a single normal population. Running the Shapiro-Wilk test on the pooled set of eight background measurements gives $SW = 0.7289$ on the original

scale and $SW = 0.8544$ after log-transforming the data. Since the critical point for the test at the $\alpha = .10$ level of significance is $sw_{.10,8} = 0.851$ (from **Table 10-3** of **Appendix D**), the results suggest that the data should be fit to a lognormal model. The log-transformed statistics for the joint background and compliance well are also found in the above table.

Step 2. Construct the prediction limit on the pooled and logged background observations. Then $n = 8$, the log-mean is 2.533, and the log-standard deviation is 0.706. Since there are 4 observations in the compliance well, take $p = 4$ as the order of the mean to be predicted. Then setting $(1-\alpha) = .99$, the Student's $t$-quantile with $(n-1) = 7$ degrees of freedom and cumulative probability of .99 is found from **Table 16-1** in **Appendix D** to be 2.998. Using equation [18.8], the upper prediction limit on the log-scale is computed as:

$$PL = 2.533 + (2.998)(0.706)\sqrt{\frac{1}{4} + \frac{1}{8}} = 3.83 \text{ log(ppb)}$$

Step 3. Compare the log-mean of the chrysene measurements at Well 3 against the upper prediction limit. Since it is less than the limit, there is insufficient evidence of chrysene contamination at this well at the $\alpha = 0.01$ significance level. ◄

## 18.3 NON-PARAMETRIC PREDICTION LIMITS

Two basic remedies are available when a data set cannot be even approximately normalized, often due to the presence of a significant fraction of non-detects. If the sample includes left-censored data (*e.g.*, non-detects), a fit to normality can be attempted using censored probability plots (**Chapter 15**) in conjunction with either the *Kaplan-Meier* or *Robust Regression on Order Statistics* [Robust ROS] techniques (**Chapter 15**). If a reasonable normality fit can be found, a parametric prediction limit can be applied. Otherwise, a non-parametric prediction limit can be considered. A non-parametric upper prediction limit is constructed by setting the limit as a large *order statistic* selected from background (*e.g.*, the maximum or second-largest background value).

As with their parametric counterparts, non-parametric prediction limits have an associated *confidence level* $(1-\alpha)$ which indicates the probability that the prediction interval $[0, PL]$ will accurately contain all $m$ of a set of $m$ future values over repeated application on many similar data sets. Unlike parametric limits, the confidence level for non-parametric limits is not *adjustable*. Despite being easily constructed for a fixed background sample size and the number of comparisons, the confidence level associated with the any maximal value used as the prediction limit is also fixed. To increase the confidence level, the primary choices are to *decrease* the number of future values to be predicted, or *increase* the number of background observations.

If existing background can be supplemented with data collected from other background wells (e.g., in interwell testing), a non-parametric test confidence level can be increased. Larger samples also provide a better characterization of site spatial variability. Unfortunately, it may always not be possible to supplement background. In these cases, another option to achieve a desired confidence level and

correspondingly control the false positive rate is to incorporate a retesting strategy as outlined in **Chapter 19**.

Although non-parametric prediction limits do not require a presumption of normality, other assumptions apply equally to both parametric and non-parametric limits. Checks should be made of statistical independence, identical distributions (under the null hypothesis), and stationarity over time and space as discussed in **Chapter 3** and **Part II** of the guidance. One particular caution for non-parametric limits is that background should ideally be screened ahead of time for possible outliers, since the upper prediction limit may be set to the background maximum or second highest observed value. Unfortunately, this often cannot be accomplished with a formal statistical test. Outlier tests are rather sensitive to the underlying distribution of the data. If this distribution cannot be adequately determined due to the presence of non-detects, an outlier test is not likely to give reliable results.

Instead of a formal test, it may be possible to screen for outliers using box plots (**Chapter 12**). Even with non-detects, the box plot 'whiskers' delineating the concentration range associated with possible outliers are computed from the sample lower and upper quartiles (*i.e.*, the 25th and 75th percentiles), which may or may not be impacted by data censoring, or perhaps mildly so when computing the lower quartile. For large fractions of non-detects, the best that can usually be done is to identify a suspected outlier through close examination of laboratory results and chain-of-custody reports.

One of two steps can be taken in the event a possible outlier is flagged. If an error has occurred, it should be corrected before constructing the prediction limit. If an error is merely suspected but cannot be proven, the prediction limit can be constructed as another order statistic from background instead of the maximum (*e.g.*, the second largest value). This will prevent the suspected outlier from being adopted as the upper prediction limit without ignoring the possibility that it may be a real measurement.

## 18.3.1 PREDICTION LIMIT FOR M FUTURE VALUES

### BACKGROUND AND PURPOSE

Given $n$ background measurements and a desired confidence level $(1-\alpha)$, a non-parametric prediction limit test for $m$ future values is an $m$-of-$m$ comparison rule. *All m* future samples need to not exceed the upper prediction limit for the test to pass. Thus the procedure is an exact parallel to the parametric prediction limit for future values. Because the method is non-parametric, no distributional model needs to be fit to the background measurements. It is assumed that the compliance point data follow the same distribution as background under the null hypothesis — even if this distribution is unknown. Although no distributional model is assumed, order statistics of any random sample follow certain probability laws which allow the statistical properties of the non-parametric prediction limit to be determined.

Once an order statistic of the sample data (*e.g.*, the maximum value) is selected as the upper prediction limit, Guttman (1970) has shown that the statistical *coverage* of the interval — that is, the fraction of the background population actually contained within the prediction interval — when constructed repeatedly over many data sets, has a *beta probability density* with cumulative distribution equal to

$$I_t(j, n-j+1) = \int_{u=0}^{t} \frac{\Gamma(n+1)}{\Gamma(n-j+1)\Gamma(j)} u^{j-1}(1-u)^{n-j}\, du \qquad [18.10]$$

where $n$ = sample size, $j$ = (rank of prediction limit value), and $\Gamma(n) = (n-1)! = (n-1)\times(n-2)\ldots\times 2\times 1$ denotes the gamma function. If the maximum is selected as the prediction limit, its rank is equal to $n$ and so $j = n$. If the second largest value is chosen as the limit, its rank would be equal to ($n$–1) and so $j$ = ($n$–1). The confidence probability for predicting that one future observation (*i.e.*, $m$ = 1) from a compliance well does not exceed the prediction limit is equal to the *expected or average coverage* of the non-parametric prediction limit.

Because of these properties, the confidence probability for a prediction limit on one future measurement can be shown to equal $(1-\alpha) = j/(n+1)$. If the background maximum is taken as the upper prediction limit, the confidence level thus becomes n/(n+1). Gibbons (1991a) has shown that the probability of having $m$ future samples all not exceed such a limit is $(1-\alpha) = n/(n+m)$. More generally, the same probability when the $j$th order statistic is taken as the upper prediction limit becomes (Davis and McNichols, 1999):

$$1-\alpha = \frac{(j+m-1)\cdot(j+m-2)\ldots\cdot(j+1)\cdot j}{(n+m)\cdot(n+m-1)\ldots\cdot(n+2)\cdot(n+1)} \qquad [18.11]$$

**Table 18-1** in **Appendix D** lists these confidence levels for various choices of $j$, $n$, and $m$. The false positive rate ($\alpha$) associated with a given prediction limit can be computed as one minus the confidence level. As this table illustrates, the penalty for not knowing the form of the underlying distribution can be severe. If a non-parametric prediction limit is to be used, *more background observations are needed compared to the parametric setting in order to construct a prediction interval with sufficiently high confidence*. As an example, to predict $m$ = 2 future samples with 95% confidence, at least 38 background samples are needed. Parametric prediction intervals do not require as many background measurements precisely because the form of the underlying distribution is assumed to be known.

It is possible to create an approximate non-parametric limit with background data containing all non-detects, by using the RL (often a quantitation limit) as the PL. A quantified value above the PL would constitute an exceedance. A superior procedure is recommended in this guidance, using the Double Quantification Rule described in **Chapter 6**.

## PROCEDURE

Step 1. Sort the background data into ascending order and set the prediction limit equal to the maximum, the second-largest observed value or another large background order statistic. Then use **Table 18-1** of **Appendix D** to determine the confidence level ($1-\alpha$) associated with predicting the next $m$ future samples.

Step 2. Compare each of the $m$ compliance point measurements to the upper prediction limit [*PL*]. Identify significant evidence of possible contamination at the compliance well if one or more measurements exceed the *PL*.

Step 3. Because the risk of false positive decision errors is greatly increased if the confidence level drops substantially below a target rate of at least 90% to 95%, the actual confidence level (as identified by equation [18.11]) needs to be routinely reported and noted whenever it is below the target level.

Note that equation [18.11] assumes the prediction limit is applied to only one constituent at a single well. If multiple tests must be conducted and a retesting procedure is employed, the confidence level of the prediction limit must be determined using the tables described in **Chapter 19**.

►EXAMPLE 18-3

Use the following trichloroethylene data to compute a non-parametric upper prediction limit for the next $m$ = 4 monthly measurements from a downgradient well and determine the level of confidence associated with the prediction limit.

| | **Trichloroethylene Concentrations (ppb)** | | | |
|---|---|---|---|---|
| | Background Wells | | | Compliance |
| Month | BW-1 | BW-2 | BW-3 | CW-4 |
| 1 | <5 | 7 | <5 | |
| 2 | <5 | 6.5 | <5 | |
| 3 | 8 | <5 | 10.5 | 7.5 |
| 4 | <5 | 6 | <5 | <5 |
| 5 | 9 | 12 | <5 | 8 |
| 6 | 10 | <5 | 9 | 14 |

SOLUTION

Step 1. Determine the background maximum and use this value to estimate the non-parametric prediction limit. In this case, the maximum value of the $n$ = 18 pooled background observations is 12 ppb. Set $PL$ = 12 ppb.

Step 2. Compare each of the downgradient measurements against the prediction limit. Since the value of 14 ppb for Month 6 exceeds PL, conclude that there is statistically significant evidence of an increase over background at CW-4.

Step 3. Compute the confidence level and false positive rate associated with the prediction limit. Since four future samples are being predicted and $n$ = 18, the confidence level equals $n/(n + m)$ = 18/22 = 82%. Consequently, the Type I error or false positive rate is at most (1 – 0.82) = 18% and the test is significant at the $\alpha$ = 0.18 level. This means there is nearly a one in five chance that the test has been falsely triggered. Only additional background data and/or use of a retesting strategy would lower the false positive rate. ◄

## 18.3.2 PREDICTION LIMIT FOR A FUTURE MEDIAN

### BACKGROUND AND PURPOSE

A prediction limit for a future median is a non-parametric alternative to a parametric prediction limit for a future mean (**Section 18.2.2**) when the sample cannot be normalized. In groundwater monitoring, the most practical application for this kind of limit is for medians of order 3 (*i.e.*, the median of three consecutive measurement values), although the same procedure could theoretically be employed for medians of any odd order (*e.g.*, 5, 7, *etc.*). The comparison rule in this case is that the test passes only if the *median* of a set of 3 compliance point measurements does not exceed the upper prediction limit. Note that this is also the same as a 2-of-3 test, whereby the well is deemed in compliance if at least 2 of 3 consecutive observations fall within the prediction interval. Therefore, only 2 independent observations will generally be needed to complete the test at uncontaminated wells. The third measurement will be irrelevant if the first two pass and so will not need to be collected.

Given $n$ background measurements and a desired confidence level $(1-\alpha)$, a non-parametric prediction limit for a future median involves a confidence probability that the median of the next $p$ future observations will not exceed the limit. As noted in **Section 18.3.1**, order statistics of any random sample follow certain probability laws. In particular, the statistical coverage ($C$) of a prediction limit estimated by the $j$th order statistic (that is, the $j$th largest value) in background will follow a ***beta distribution*** with parameters $j$ and $(n+1-j)$. Following the notation of Davis and McNichols (1987), the ***conditional probability*** that the median of 3 independent future values will not exceed the non-parametric prediction limit can be shown to equal

$$\Pr\left\{\text{Future median inbounds} \,\middle|\, X_{j:n}\right\} = 3C^2 - 2C^3 \qquad [18.12]$$

where $X_{j:n}$ denotes that the prediction limit equals the $j$th largest order statistic in a sample of $n$ observations and a conditional probability denotes the chance that an event will occur given the observance of another event (in this case, after having observed $X_{j:n}$). The (unconditional) confidence probability $(1-\alpha)$ can then be derived by taking the ***expected value*** of equation [18.12] with respect to the random variable $C$. Using standard properties of the beta distribution, this probability becomes:

$$1-\alpha = \frac{(3n-2j+5)(j+1)j}{(n+3)(n+2)(n+1)} \qquad [18.13]$$

Thus the confidence level associated with a prediction limit for a future median of order 3 depends simply on the sample size of background ($n$) and the order statistic selected as the upper prediction limit ($j$). **Table 18-2** in **Appendix D** provides values of the confidence level for various $n$ and choices of the order statistic. Like the non-parametric prediction limit for $m$ future values, ease of construction comes with a price. More background measurements are required to achieve the same levels of confidence attainable via a parametric prediction limit for a future mean. For instance, to achieve 99% confidence in predicting a median of order 3 in a single test, at least 22 background observations are needed if the maximum is selected as the upper prediction limit, and at least 40 background observations are needed if the prediction limit is set to the second largest measurement. Parametric prediction intervals do not

require as many background samples precisely because the form of the underlying distribution is assumed to be known.

### REQUIREMENTS AND ASSUMPTIONS

Once an order statistic (of rank *j*) is selected as the upper prediction limit, the confidence level is fixed by the number of background samples (*n*). The confidence level can only be increased by enlarging background. However, equation [18.13] is only applicable for the case of predicting a future median of *a single constituent at a single well*. To account for multiple tests and to incorporate a retesting strategy (both of which are usually needed), the specific strategies and tables of confidence levels presented in **Chapter 19** should be consulted.

### PROCEDURE

Step 1. Sort the background data into ascending order and set the upper prediction limit [*PL*] equal to one of the following: the background maximum, the second largest value, or another large order statistic in background. If the largest background measurement is a non-detect, set an approximate upper prediction limit as the RL most appropriate to the data (usually the lowest achievable quantitation limit [QL]).

Step 2. Compute the median of the next three consecutive compliance point measurements. Compare this statistic to the upper prediction limit. Identify significant evidence of possible contamination at the compliance well if the median exceeds *PL*. If *PL* equals the RL, identify an exceedance, if the median is quantified above the reporting limit.

Step 3. Based on the background sample size (*n*), use **Table 18-2** of **Appendix D** to determine the confidence level ($1-\alpha$) associated with predicting the median of the next $p = 3$ future measurements. Because the risk of false positive errors is greatly increased if the confidence level drops much below a targeted rate of at least 90% to 95%, the actual confidence level (as identified in equation [18.13]) should be routinely reported and noted whenever it is below the target level.

Note that equation [18.13] assumes the prediction limit is applied to only one constituent at a single well. If multiple tests are conducted and a retesting procedure is employed, the confidence level of the prediction limit needs to be determined using the tables described in **Chapter 19**.

### ►EXAMPLE 18-4

Use the following xylene background data to establish a non-parametric upper prediction limit for a future median of order 3. Then determine if the compliance well shows evidence of excessive xylene contamination.

| | Xylene Concentrations (ppb) | | | |
|---|---|---|---|---|
| | | Background | | Compliance |
| Month | Well 1 | Well 2 | Well 3 | Well 4 |
| 1 | <5 | 9.2 | <5 | |
| 2 | <5 | <5 | 5.4 | |
| 3 | 7.5 | <5 | 6.7 | |
| 4 | <5 | 6.1 | <5 | |
| 5 | <5 | 8.0 | <5 | |
| 6 | <5 | 5.9 | <5 | <5 |
| 7 | 6.4 | <5 | <5 | 7.8 |
| 8 | 6.0 | <5 | <5 | 10.4 |

SOLUTION

Step 1. The maximum value in the set of pooled background measurements is 9.2. Assign this value as the non-parametric upper prediction limit, $PL = 9.2$.

Step 2. Compute the median of the three compliance measurements. This statistic equals 7.8 ppb. Since the median does not exceed PL, there is insufficient evidence of xylene contamination at Well 4, despite the fact that the *maximum* at Well 4 is larger than the maximum observed in background.

Step 3. Compute the confidence level and false positive rate associated with this prediction limit. Given that $n = 24$ and the order statistic selected is the maximum (*i.e.*, $j = n$), use **Table 18-2** in **Appendix D** to determine that the confidence level for predicting a future median of order 3 equals 99.1% and therefore the Type I error or false positive rate is at most 0.9%. ◄

# CHAPTER 19. PREDICTION LIMIT STRATEGIES WITH RETESTING



This chapter is a core part of the recommended statistical approach to detection monitoring. Even the smallest of facilities will perform enough statistical tests on an annual basis to justify use of a retesting strategy. Such strategies are described in detail in this chapter in conjunction with prediction limits. First, the Unified Guidance considers the concept and computation of site-wide false positive rates [SWFPR]. Then different retesting strategies useful for groundwater monitoring are presented, including:

- Parametric prediction limits with retesting (**Section 19.3**), and
- Non-parametric prediction limits with retesting (**Section 19.4**)

## 19.1 RETESTING STRATEGIES

*Retesting* is a statistical strategy designed to efficiently solve the problem of *multiple comparisons* (*i.e.*, multiple, simultaneous statistical tests). An introduction to multiple comparisons is presented in **Chapter 6**. At first glance, formal retesting seems little different than a repackaged form of *verification resampling*, a practical technique used for years to double-check or verify the results of initial groundwater sampling. Indeed, all retesting schemes are predicated on the idea that when the initial groundwater results indicate the presence of potentially contaminated groundwater, one or more additional groundwater samples should be collected and tested to determine whether or not the first results were accurate.

The difference between formal retesting schemes and verification resampling found in the regulations is that the former *explicitly incorporates the resample(s) into the calculation of the statistical properties of the overall test*. A statistical "test" then needs to be redefined to include not only the statistical manipulation of the initial groundwater sampling results, but also that for any further resamples. Both the initial samples and the resamples are integral components of any retesting method.

The principal advantage of retesting is that very large monitoring networks can be statistically tested without necessarily sacrificing either an acceptable false positive rate or adequately high *effective power*. Data requirements for a typical retesting scheme are often less onerous than those required for an analysis of variance (ANOVA). Instead of having to sample each well perhaps four times during any

given evaluation period, many of the retesting strategies discussed below involve a minimum of one new sample at each compliance well. Resamples are collected only at wells where the initial results exceed a limit, and no explicit *post-hoc* testing of individual wells is necessary as with ANOVA in order to identify a contaminated well.

Since a statistical test utilizing retesting is not complete until all necessary resamples have been evaluated, it is important to outline the formal *decision rules* or *scheme* associated with each retesting strategy. Retesting schemes presented in the Unified Guidance fall into two types: 1-of-*m* and the modified California approach. The 1-of-*m* approach was initially suggested by Davis and McNichols (1987) as part of a broader method termed "*p*-of-*m*." The 1-of-*m* scheme assumes that as many as *m* samples might be collected for a particular constituent at a given well, including the initial groundwater sample and up to (*m*–1) resamples.

1-of-*m* schemes are particularly attractive as retesting strategies. If the initial groundwater observation is in-bounds, the test is complete and no resamples need to be collected. Only when the first value exceeds the background prediction limit, does additional sampling come into play. For practical reasons, only 1-of-*m* schemes with *m* no greater than 4 are considered in the Unified Guidance. A 1-of-4 retesting plan implies that *up to* 4 groundwater measurements may have to be collected at each compliance well, including the initial observation and 3 possible resamples. For the test to be valid, all of these sample measurements need to be statistically independent. This generally requires that sufficient time elapses *between resample collection* so that the assumption of statistical independence or lack of autocorrelation is reasonable (see the discussion in **Chapter 14**). Because many groundwater evaluations are conducted on a semi-annual basis, three will generally be a practical upper bound on the number of independent resamples that might be collected. Thus the 1-of-2, 1-of-3, and 1-of-4 retesting schemes are included below.

The second type of retesting scheme is known as the modified California approach. The decision rules for this test are slightly different from the 1-of-*m* schemes, although the test passes as before if the initial groundwater measurement is inbounds. If it exceeds the background limit, two of the three resample need to be inbounds for the test to pass. The modified California strategy thus requires a *majority* of the resamples to be inbounds for a compliance well test to be deemed 'in bounds'. A 1-of-4 scheme could have both the initial value and the first two resamples be out-of-bounds, yet pass the test with an inbounds result from the third resample. Although the modified California test appears to be more stringent, the prediction limit for a 1-of-4 test under the same input conditions will be lower and hence be more likely to trigger single comparison exceedances. With the prediction limits correctly defined, both will have identical false positive errors for any specific monitoring design. The guidance also provides the same four non-parametric versions of the 1-of-*m* and modified California tests for future values.

A useful variation on the 1-of-*m* retesting scheme for individual measurements is the 1-of-*m* strategy for *means or medians*. Instead of testing a series of individual values, a series of means or medians of order *p* is tested. The *order* of the mean or median refers to the number of individual measurements used to compute the statistic. For example, 1-of-2 retesting with means of order 2 requires that a pair of initial observations be averaged and the resulting mean compared against the background limit. If that initial mean is out-of-bounds, a second pair of observations (*i.e.*, two resamples) would be

collected and averaged to form the resample mean. The test would fail only if both the initial mean and the resample mean exceeded the background limit.

Retesting schemes for means or medians have steeper data requirements than retesting strategies for individual measurements and may not be practical at many sites. Nevertheless, the statistical properties (*e.g.*, power and false positive rate) associated with the testing of means and medians are superior to comparable plans on individual observations. The Unified Guidance provides five mean retesting plans: 1-of-1, 1-of-2, or 1-of-3 for means of order 2; and 1-of-1 and 1-of-2 for means of order 3. The guidance also provides 1-of-1 and 1-of-2 tests of medians of size 3 as non-parametric options.

These plans were chosen to limit the maximum possible number of distinct and independent sampling measurements per compliance well during a single evaluation period to six. In fact, the data requirements vary substantially by scheme. With means of order 2, the 1-of-1 plan requires a maximum of two new sample measurements; the 1-of-2 plan requires as many as four; while only the 1-of-3 plan might need a total of six. For means of order 3, the 1-of-1 plan requires three new measurements to form the single mean; the 1-of-2 plan might require up to six. But for higher order 1-of-*m* mean or median tests, only the initial samples may be needed to identify a 'passing' test outcome under most background conditions.

The three 1-of-1 mean and median plans provided in the guidance are technically not retesting schemes. The decision rule for these plans merely requires a comparison of a single mean or median against the background limit. If the initial mean or median comparison is inbounds, the test passes. If not, the test fails. The fact that each average is computed from multiple individual measurements implies that an implicit retest or verification resampling is built into these strategies. The statistical properties of the 1-of-1 plans can often be better than comparable 1-of-*m* schemes for individual values, with fairly similar sampling requirements.

The Unified Guidance provides 1-of-1 and 1-of-2 non-parametric prediction limit tests for future medians of order 3. By ‘median of order 3’, it means that the median or ‘middle value’ of a set of three consecutive sampling events. In the 1-of-2 case, the test passes if either the initial median is inbounds or, if not, when the resample median is inbounds. The 1-of-1 scheme does not involve any resampling, but does require at least two distinct sampling measurements to determine whether the initial median is inbounds.[1]

As discussed in **Chapter 6**, proper design of a groundwater detection monitoring program will generally require an initial choice of a retesting scheme ***before*** future or compliance sampling data have been collected. As a practical matter, sample collection should be spaced far enough apart in time to ensure that any potentially needed resamples are statistically independent. Thus, the maximum number of resamples need to be known in advance in order to structure a feasible sampling plan for a particular retesting strategy. Each retesting scheme also involves a different set of decision rules for evaluating the status of any given compliance well. The rules will determine how the background limit will be computed. Given the same background sample and group of compliance wells, different retesting schemes lead to ***different*** background limits on the ***same*** data.

[1] As noted in **Chapter 18**, the 1-of-1 retesting scheme for medians of order 3 is equivalent as a decision rule to a 2-of-3 scheme for individual measurements.

If parametric prediction limits are used, the general formula for the limit introduced in **Chapter 18** is $\bar{x} + \kappa s$. The $\kappa$-multiplier and thus the prediction limit will vary depending on which 1-of-*m* or modified California plan is chosen. The $\kappa$-multipliers also depend on the monitoring evaluation schedule in place at the facility. In typical applications, it is expected that the background sample used in statistical evaluations from any given year will either be *static* or substantially overlap from one evaluation to the next. The same background observations are likely to be utilized or will substantially overlap if newer background data are added to the existing pool. Since at least a subset of the background measurements will be commonly employed in all the evaluations, there will be a *statistical dependence* exhibited between distinct evaluations (see **Section 19.2** below). The number of evaluations per year against a common background will affect the correct identification of prediction limits. Consequently, the evaluation schedule (*i.e.*, annual, semi-annual, quarterly) also needs to be known or specified in advance.[2]

## 19.2 COMPUTING SITE-WIDE FALSE POSITIVE RATES [SWFPR]

As discussed in **Chapter 6**, the fundamental purpose of detection monitoring is to accurately identify a significant change in groundwater relative to background conditions. To meet this objective, statistical monitoring programs should be designed with the twin goals of ensuring adequate statistical power to flag well-constituent pairs elevated above background levels and limiting the risk of *falsely* flagging uncontaminated wells across an entire facility. The latter is accomplished by addressing the *site-wide false positive rate* [SWFPR]. Both goals contribute to accurate evaluation of groundwater and to the validity of statistical groundwater monitoring programs.

Retesting significantly aids this process of meeting *both* criteria. However, it can be much easier to design and implement an appropriate retesting scheme if one understands how the SWFPR is derived. The SWFPR is based on the assumptions that no contamination is actually present at on-site monitoring wells, and that each well-constituent pair in the network behaves independently of other constituents and wells from a statistical standpoint. If $Q$ denotes the probability that a particular well-constituent pair will be falsely declared an exceedance (a *false positive* error), the probability of at least one such false positive error among $r$ independent tests is given by:

$$\alpha = SWFPR = 1 - \left(1 - Q\right)^r \qquad [19.1]$$

$(1–Q)$ equals the chance that the test will correctly identify the well-constituent pair as 'inbounds.' The value of $Q$ itself will depend on the type of retesting scheme being used.

[2] The Unified Guidance distinguishes between the statistical *evaluation* (or testing) schedule and the *sampling* schedule. Regularly scheduled sampling events might occur quarterly, even though a statistical evaluation of the data only occurs semi-annually or annually. Further, resamples do not constitute regular sampling events, since they are only collected at wells with initial exceedances, but they *are* associated with the data for a particular evaluation. By separately identifying the evaluation schedule, there is 1) less confusion about the role of resamples in the testing process, and 2) opportunity to design monitoring programs, so as to allow for multiple individual observations to be collected prior to each evaluation.

Consider a 1-of-3 retesting plan for future observations. A false positive at a given well-constituent pair will be registered only if all three observations — the initial groundwater measurement and two resamples — exceed the background prediction or control limit. If ω represents the probability that one of these observations exceeds the background limit, *Q* can be calculated as $\omega \times \omega \times \omega$ (since the initial measurement and resamples are statistically independent) and the SWFPR as:

$$\alpha = SWFPR = 1 - \left(1 - \omega^3\right) \qquad [19.2]$$

By setting the target site-wide α equal to 0.10 and solving for ω, one could potentially compute the individual comparison false positive rate ($\alpha_{comp} = \omega$) associated with any single comparison against the background limit. This would identify the individual *per-comparison* confidence level ($1 - \alpha_{comp}$) necessary to compute the background limit in the first place.[3] If the background limit is computed as a prediction limit for the next single future measurement (*i.e.*, *m* = 1 in a 1-of-*m* scheme), then ω equals the probability that a single new observation (independent of background) exceeds the prediction limit, and (1–ω) equals the confidence level of that prediction limit. Further, since ω can be obtained from equation [19.2] as:

$$\omega = \sqrt[3]{1 - \left(1 - \alpha\right)^{1/r}} \qquad [19.3]$$

the upper prediction limit for a site involving 500 tests (for instance, 50 wells and 10 constituents per well) and 20 background samples could be computed using an individual, per-comparison confidence level of

$$1 - \omega = 1 - \sqrt[3]{1 - \left(1 - .10\right)^{1/500}} = 1 - .0595 = 94.0\%$$

leading to a final prediction limit of

$$PL = \bar{x} + t_{.94,19} s \sqrt{1 + \frac{1}{20}}$$

where $\bar{x}$ and *s* are the background sample mean and standard deviation.

Unfortunately, certain statistical dependencies render the foregoing calculations somewhat inaccurate. Whether or not a resample exceeds the background limit for any constituent depends partly on whether the initial observation for that test *also* eclipsed the limit. This is because the ***same background data*** are used in the comparison of both the initial groundwater measurement and the resamples. This creates a statistical dependence between the ***comparisons,*** even when the compliance point observations themselves are statistically ***independent***. If the background data sample mean happens to be low relative to the true population mean, the background limit will tend to be low. Each of the compliance point observations (whether the first measurement or subsequent resamples) will have a

[3] Note that $\alpha_{comp}$ does not represent the false positive rate for the complete 1-of-3 test, but is being treated for the sake of argument as a one of a series of 3 individual and independent tests.

greater than expected chance of exceeding it. Likewise, if the background sample mean is substantially higher than the population mean, the background limit will tend to be high, resulting in a lower-than-expected chance of exceedance for each of the compliance measurements.

A similar dependence occurs for each well-constituent pair tested against a single background across evaluation periods (see discussions in **Chapter 5** and **Section 19.1**). A further dependence occurs when well-constituent pairs from many compliance wells are compared to a common interwell background. The tests during each statistical evaluation again share a common (or nearly common) background, thus impacting the individual test false positive rate ($\alpha_{test}$) and the SWFPR ($\alpha$) in turn. Three common evaluation strategies are considered in the Unified Guidance: quarterly, semi-annual, and annual. The SWFPR is computed on a cumulative, annual basis, with the assumption that background and the associated background limit will not be updated or recomputed (especially for intrawell tests) more often than every one to two years.[4]

These dependencies between successive comparisons and tests against the background limit during retesting means that the derivation above will generally *not* result in a background limit with the targeted annual SWFPR of 10%. The actual false positive rate ($\alpha$) will be somewhat higher and can be substantially higher if the background sample size ($n$) is small to moderate (say less than 50 samples). In part, this is because the correlation between successive comparisons against a common background limit is on the order of $1/(1+n)$. That is, the smaller the background size, the greater the correlation between the resamples and test comparisons. The impact on the SWFPR is also greater if this dependence is ignored.

Fortunately, as Gibbons (1994) has noted, the solution suggested in the previous example will be approximately valid for large background data sets (say $n > 50$), since then the correlation between successive resamples and/or tests is minimal. In fact, an approximate solution for the modified California and more general 1-of-$m$ retesting schemes can also be derived. In the case of 1-of-$m$ schemes, the probability $Q$ of a false positive (for $m$ = 1 to 4) is $\omega^m$, leading to a SWFPR of :

$$\alpha = SWFPR = 1 - \left(1 - \omega^m\right)^r \qquad [19.4]$$

Solving for $\omega$ in equation [19.4] leads to an approximate individual comparison false positive rate ($\alpha_{comp} = \omega$) of:

$$\omega = \sqrt[m]{1 - \left(1 - \alpha\right)^{1/r}} \qquad [19.5]$$

For the modified California plan, a false positive for a given well-constituent pair during a single evaluation will be registered only if both the initial measurement and at least two of three resamples are

[4] Even with these assumptions, not all the statistical dependence will be accounted for at every site or for all constituents. Even when background is updated with new measurements, some of the already existing background values are likely to be used in re-computing the background limit. Some well-constituent pairs may be correlated, contradicting the assumption of independence between tests at the same well or for the same constituent at different wells. The Unified Guidance also does not presume to compute the SWFPR for other multi-year periods or for the life of the facility.

out-of-bounds (*i.e.*, exceed the background limit). Consequently, the probability *Q* of a false positive for that pair may be expressed as:

$$Q = \omega\left[3\omega^2\left(1-\omega\right)+\omega^3\right] = \omega^3\left(4-3\omega\right) \qquad [19.6]$$

As before, ω represents the probability of any single observation exceeding the background limit. Both the initial and any resample comparisons against the limit are assumed to be statistically independent. Given *Q*, the approximate overall false positive rate then becomes:

$$\alpha = SWFPR = 1-\left[1-\omega^3\left(4-3\omega\right)\right]^r \qquad [19.7]$$

Since ω will always be small in practice, one can usually ignore the term $\omega^4$ when expanding the right-hand side of equation [19.7]. Then the approximate SWFPR becomes:

$$\alpha \approx 1-\left[1-4\omega^3\right]^r \qquad [19.8]$$

Leading to a solution for ω:

$$\omega \approx \sqrt[3]{1-\left(1-\alpha\right)^{1/r}}\sqrt[3]{\frac{1}{4}} \qquad [19.9]$$

which can again be used to construct a background limit for a single new observation.

As an example, if the target SWFPR is 10% and one must test $r$ = 200 comparisons using the modified California plan, ω would become:

$$\omega \approx \sqrt[3]{1-.90^{1/200}}\sqrt[3]{\frac{1}{4}} = .0508 = 5.1\%$$

If the background limit is a prediction limit for the next future value, a confidence level of approximately 94.9% would be needed to achieve the desired overall false positive rate of 10%. This assumes that the background sample size is sufficiently large (say $n$ > 50) to make the correlation between retests negligible. In similar fashion, the respective single comparison error rates for the 1-of-2 through 1-of-4 tests of future observations in this example would respectively be: ω = .0229, .0808, and .1515.

### 19.2.1 BASIC SUBDIVISION PRINCIPLE

The previous section highlighted certain dependencies in statistical tests due to comparisons of one or more samples or sample sets against a common background. In the sitewide design of a facility detection monitoring system, the overall target design SWFPR is proportionately divided among all relevant tests conducted in an annual period. Depending on the type of testing (e.g., interwell versus

intrawell, or a parametric versus non-parametric), the target error rates for a portion of the total set of potential tests may need to be calculated.

Identifying false positive target rates is important when considering non-parametric prediction limit tests. The cumulative target error rate for a group of annual tests against a single constituent is needed to compare with the achievable levels in **Tables 19-19 through 19-24** in **Appendix D**. The latter achievable rates take into account the dependencies previously discussed. ***κ*-**multiple **Tables 19-1 to 19-18** in **Appendix D** for parametric prediction limit tests have already made use of target false positive rate calculations which are generally not needed for identifying the appropriate multipliers. The various dependencies against a common background are accounted for in the ***κ*-**multiple tables to meet the nominal target rates. **R-**script software for certain parametric prediction limit tests discussed in a following section and in **Appendix C** also makes use of a target per-test false positive error rate as input.

In assigning target rates, the Unified Guidance uses a basic ***subdivision principle*** which makes certain assumptions. First and foremost, it is assumed that the total suite of tests can be subdivided into mutually exclusive, independent[5] tests. Each relevant annual statistical test is assigned the same single test error rate ($\alpha_{test}$). Using the properties of the Binomial distribution, the target single test error rate can be obtained using equation [19.10] for $r$ total annual tests. The total number of annual tests $r$ is the product of the number of compliance wells ($w$), the number of valid constituents ($c$), and the number of evaluations per year ($n_E$) or $r = w \times c \times n_E$, with $\alpha$ = SWFPR:

$$\alpha_{test} = 1 - (1-\alpha)^{1/r} \qquad [19.10]$$

Then a cumulative false positive rate can be assessed for any appropriate subset of tests. This principle would apply, for instance, if there is more than one regulated unit at a site and each regulated unit can be treated independently. A consistent portion of the overall targeted false positive rate α would be assigned to each regulated unit ($\alpha_{unit}$), using a rearrangement of equation [19.10]. If a facility with three units B, C, and D had 120 total annual tests ($b + c + d = 120 = r$), the cumulative target error rate for Unit B would be: $a_{UnitB} = 1 - (1-\alpha_{test})^b$ and similarly for Units C and D. These three cumulative error rates will *approximately* (but not exactly) sum to a total sitewide value close to the SWFPR. However, as joint independent tests taken together, the annual SWFPR is in fact exactly 10%. The Bonferroni assumption makes use of the approximately linearity of such error rates for SWFPR calculations (discussed below).

The ways in which the overall SWFPR might be partitioned will vary with each site, considering units, types of tests, number of wells, constituents and evaluations per year. If unit-specific cumulative false positive rates were established, the group of tests associated with each monitoring constituent within each unit might be separately considered. Each group might potentially be further subdivided into intrawell versus interwell tests, or prediction limits versus control charts, *etc.*, assuming a mixture of statistical methods is employed. By using the subdivision principle in a consistent way, the targeted SWFPR can be accurately maintained.

[5] The Unified Guidance does not presume that every statistical test is in fact independent. Tests or groups are treated as if independent, however, to allow the computation of nominal target false positive rates and/or to be consistent with regulatory constraints (*e.g.*, all constituents must be tested separately).

One important use when calculating SWFPR rates is to account for multiple constituents. In particular, non-parametric test theory is applied to only a single constituent at a time. Since each constituent has its own set of background data and presuming the constituents behave independently of one another, the dependence caused by using a common background pertains only to those comparisons made against the background for that constituent. To clarify this concept, suppose a total set of $r$ tests consists of $c$ separate chemicals each monitored at $w$ wells annually (*i.e.*, $r = c \times w \times n_E$ and $n_E = 1$). For each constituent, the dependence caused by a common background only applies to the $w$ comparisons (one for each well) made for that monitoring parameter. This means that the overall target $\alpha$ = SWFPR needs to be apportioned into a fraction for each constituent, called the per-constituent false positive rate or $\alpha_c$. This can be done using the Binomial formula based on the single test error rate for $w$ wells as: $a_c = 1 - (1-\alpha_{test})^{w \cdot n_E}$ or by partitioning the overall $\alpha$ to each constituent $c$:

$$\alpha_c = 1 - (1-\alpha)^{1/c}$$

The two calculations are equivalent under these conditions, with the latter equation somewhat easier to use.

A similar situation occurs at sites requiring a combination of interwell and intrawell tests. Computation of the SWFPR can be appropriately handled using the basic subdivision principle. For interwell tests, measurements collected at each compliance well are compared against a common interwell background, creating a degree of statistical dependence not only between successive individual test comparisons (*i.e.*, initial sample and any resamples) at a given well, but also between tests at different compliance wells. With intrawell tests, each well supplies its own background. This implies that the component of between-well test dependence is eliminated, changing the way κ-multipliers for intrawell background limits with retesting are computed.

For a given set of $r$ well-constituent pairs, $l$ tests to be conducted on an interwell basis, and the remaining $(r - l)$ tests conducted as intrawell, two cumulative false positive rates need to be computed. The single test false positive error rate $\alpha_{test}$ approach can be used: $a_{\text{int } er} = 1 - (1-\alpha_{test})^{l}$ for the subset of $l$ interwell tests, and $a_{\text{int } ra} = 1 - (1-\alpha_{test})^{r-l}$ for the subset of $r - l$ intrawell tests, in order to correctly maintain the SWFPR equal to $\alpha$. A somewhat more direct approach can also be used: $\alpha_{\text{int } er} = 1 - (1-\alpha)^{l/r}$ for the interwell tests and $\alpha_{\text{int } ra} = 1 - (1-\alpha)^{(r-l)/r}$ for the intrawell tests. The two sets of equations are consistent.

In general, the subdivision principle works as follows. If a group of $r$ tests with targeted false positive rate, $\alpha$, is divided into $s$ distinct and mutually exclusive independent subsets, the false positive rate for each subset ($\alpha_{sub}$) can be computed as:

$$\alpha_{sub} = 1 - (1-\alpha)^{1/s} \qquad [19.11]$$

The basic subdivision principle does not guarantee that the resulting detection monitoring program will have sufficient effective power to match the EPA reference power curve (ERPC). The foregoing calculations merely point to the correct overall false positive rate.

As discussed in Section 6.2.2 of **Chapter 6**, a simpler approach would be to partition the overall SWFPR among a facility's annual number of tests, and can make use of the Bonferroni approximation. With low false positive rates characteristic of detection monitoring design, the total SWFPR can be divided by the number of annual tests for any of the various combinations of constituents, separate units, or interwell versus intrawell tests. The Bonferroni approach results in slightly different false positive values than by directly using the Binomial formula, as described above.

As an overall example, assume a facility with $w$ = 20 wells monitored twice per year ($n_E$ = 2) for $c$ = 8 constituents Further, assume that 5 of the constituents can be monitored interwell and 3 need to be handled as intrawell comparisons. Non-parametric prediction limits will be considered for all tests. Calculate the target cumulative false positive error rates for interwell and intrawell comparisons, with the SWFPR = $\alpha$ = .1.

This site has a total of $r = w \times c \times n_E$ = 20×8×2 = 320 tests per year. For the five interwell constituents, there are 20×2×5 = 200 tests, with 20×2×3 = 120 intrawell tests. Each of the 5 interwell constituents will have 20×2 = 40 tests against a common background, while 2 semi-annual sample tests will be made against each of the 20×3 = 60 intrawell backgrounds.

From equation [19.10], the single test false positive error rate is: $\alpha_{test} = 1-(1-\alpha)^{1/r}$ = $1-(1-.1)^{1/320} = .0003292$. Each set of interwell constituent tests will have a cumulative false positive error rate $\alpha_c$ for the 40 annual tests as: $\alpha_c = 1-(1-\alpha)^{1/c} = 1-(1-.1)^{1/8} = .01308$. Note that *all 8* constituents are used in the equation, since the same false positive error rate is uniformly applied to all distinct subgroup tests. The result can be obtained using the single test error rate equation: $a_c = 1-(1-\alpha_{test})^{w \cdot n_E} = 1-(1-.0003292)^{40} = .01308$. This target value would be used to compare with achievable non-parametric test error rates for the same input conditions. The cumulative interwell error rate for all five constituents can be calculated as: $a_{\text{int } er} = 1-(1-\alpha_c)^c = 1-(1-.01308)^5 = .06371$.

For the intrawell tests, the simplest approach uses the single test error rate for two tests: $a_{2-\text{int } ra} = 1-(1-\alpha_{test})^{w \cdot n_E} = 1-(1-.0003292)^{1 \cdot 2} = .0006583$. This would be the cumulative error rate to consider with non-parametric intrawell tests. The overall intrawell cumulative error rate for the sixty tests would then be: $a_{60-\text{int } ra} = 1-(1-\alpha_{2-\text{int } ra})^{w \cdot c} = 1-(1-.0006583)^{60} = .03873$.

If the two overall interwell and intrawell cumulative error rates were added, the sum is .1024, quite close to the nominal 10% SWFPR. It is exactly that value when considered jointly. By comparison the single test error rate using the Bonferroni approximation would be .1/320 = .0003125, while the exact Binomial value is .0003292. The estimated interwell cumulative error for a single constituent would be 40 times the single test value or .0125 (versus the calculated .01308). For many non-parametric test considerations, these differences are relatively minor.

# 19.3 PARAMETRIC PREDICTION LIMITS WITH RETESTING

## BACKGROUND AND PURPOSE

Upper prediction limits for *m* future observations and for future means were described in **Chapter 18**. Applied to a network of statistical comparisons in detection monitoring, these procedures can be considered an extension to Dunnett's *multiple comparison with control* [MCC] procedure (Dunnett, 1955). These procedures explicitly incorporate retesting that is applicable to a wider variety of cases than addressed by Dunnett.

Retesting can be incorporated with either interwell or intrawell prediction limits. Depending on which approach is adopted, there is a distinct difference in the κ-multipliers of the general prediction limit formula. In an interwell retesting strategy, there are at least two forms of statistical dependence that impact the SWFPR. One is that each initial measurement or resample at a given compliance well is compared against the same background. A second is the dependence among compliance wells and number of annual evaluations, all of which are compared against a common upgradient background. In intrawell retesting, this second form of dependence is either essentially eliminated if there is only one annual statistical evaluation or else substantially reduced in the event of multiple evaluations.[6] The remaining dependence is among successive resamples at each well.

To account for the basic differences between interwell and intrawell prediction limit tests, an extensive series of tables is provided in **Appendix D** listing a wide combination of background sample sizes, numbers of wells, numbers of constituents, and distinctions between interwell and intrawell tests. In conjunction with an evaluation schedule (*i.e.*, annual, semi-annual, or quarterly), these tables can be used to design and implement specific parametric retesting strategies in this chapter. All of the κ-multiplier tables for parametric prediction limits are structured to meet an annual SWFPR of 10% per year and to accommodate groundwater networks ranging in size from one to 8,000 total statistical tests per year. The Unified Guidance tables are more extensive than similar tables in Gibbons (1994b). Further, each table is designed to indicate the effective power of the $\kappa$-multiplier entries.

If a particular network configuration is not directly covered in the **Appendix D** tables, two basic options are available. First, bilinear interpolation can be used to derive an approximate $\kappa$-multiplier (see below for guidance on table interpolation). Second, the free-of-charge, open source, and widely available **R** statistical programming package (www.r-project.org) can be employed to compute an exact $\kappa$-multiplier. Further instructions and the two template codes used to compute the Unified Guidance κ-multiplier tables are provided in **Appendix C**. After installing the **R** package, these template codes can be run by supplying specific parameters for the network of interest (*e.g.*, number of wells, constituents, background sample size, *etc.*). Some familiarity with properly installing a program like **R** is helpful. **Appendix C** explains how to execute a pre-batched set of commands. No other technical programming experience is needed.

[6] If multiple evaluations occur each year, new compliance samples each evaluation period are tested against the common intrawell background.

## REQUIREMENTS AND ASSUMPTIONS

The basic assumptions of parametric prediction limits were described in **Chapter 18**. These include data that are normal or can be normalized (via a transformation), lack of outliers, homogeneity of variance between the background and compliance point populations, absence of trends over time, stationarity, and statistical independence of the observations.

The Unified Guidance provides *separate* $\kappa$-tables for interwell and intrawell limits. One of these approaches should be justified before computing prediction limits. To use *interwell* prediction limits, there should be no significant natural spatial variation among the mean concentrations at different well locations. Otherwise, a prediction limit test could give meaningless results, since average downgradient levels might naturally be higher than background even in the absence of a contaminant release. The assumption of spatial variability should therefore be checked using the methods in **Chapter 13**.

While *intrawell* testing eliminates the problem of natural spatial variability, intrawell background often is developed using the first $n$ samples from each compliance point well. Since historical data from compliance wells need to be utilized to do this, these groundwater measurements should be uncontaminated. The number of intrawell background samples available may also be rather limited. $n$ will tend to be initially small prior to any updating of background. Such constraints will limit the intrawell retesting schemes that can both minimize the SWFPR yet maintain effective power similar to the ERPCs.

One possible way to overcome this limitation is to estimate a *pooled standard deviation* across many wells along the lines suggested by Davis (1998). Such a calculation is no more difficult than a one-way ANOVA (**Chapter 13**) for identifying on-site spatial variability. The *mean squared error* [MSE] component of the $F$-statistic in ANOVA gives an estimate of the average per-well variability. To the extent that mean levels vary by well location ***but the population standard deviation does not***, a one-way ANOVA can be run on a collection of wells (both background and compliance) to estimate the average within-well variance, and hence, the common *intrawell* standard deviation (see **Chapter 13** for further details and examples).

Instead of a standard deviation estimate based solely on intrawell background at a single well with its attendant limits in size and degrees of freedom, the *mean* concentration level can be estimated on a well-specific basis, while the *standard deviation* is estimated utilizing a collection of wells leading to much larger degrees of freedom. Although the intrawell background size for a given well might be small (*e.g.*, $n$ = 4 or 8), the $\kappa$-multiplier used to construct the prediction limit is based on both the *effective sample size* (*i.e.*, degrees of freedom plus one) *and* the intrawell sample size (*n).*

The pooled standard deviation for intrawell comparisons can be utilized if the population standard deviation is *approximately constant across wells*. Many data sets may not appear so initially; however, any transformation to normality must first be taken into account. The standard deviation is only assumed to be constant *on the transformed scale.* Furthermore, once any transformation is applied, the collection of wells should explicitly be tested for homogeneity of variance using the tools in **Chapter 11**. *Only if the assumption of equal variances across wells seems reasonable should the pooled standard deviation estimate be used.*

With little or no spatial variability among well locations, an interwell test might be considered. However, the sample standard deviation (*s*) computed from background may not adequately estimate true background variability. This can happen when there is a temporal component to the variability affecting all wells at a site or regulated unit in parallel fashion, or when there is a significant degree of autocorrelation between successive samples.

A random, temporal component to the variability can result from changes to the laboratory analytical method or field sampling methodology, periodic re-calibration of lab instruments, or other sample handling or preparation artifacts that tend to impact all observations collected during a given sampling event. Such a temporal component can sometimes be identified through the use of parallel time series plots (**Section 14.2.1**) or through a one-way ANOVA using time-of-sampling as the factor (**Section 14.2.2**). Results of the ANOVA can be used to derive a better estimate of the background population standard deviation ($\sigma$), along with adjusted degrees of freedom for use in constructing the upper prediction limit (see **Chapter 14** for further details and an example).

When autocorrelation is present, methods to adjust the standard deviation estimate and degrees of freedom entail possibly modeling the autocorrelation function. This issue is beyond the scope of the Unified Guidance and consultation with a professional statistician is recommended. The most practical way to avoid significant autocorrelation between samples is to allow enough time to lapse between sampling events. Precisely how much time will vary from site to site, but Gibbons (1994a) and others (for instance, American Society for Testing and Materials, 2005) recommend that the frequency of sampling be no more frequent than quarterly. Alternatively, a pilot study can be run on two or three wells with the sample autocorrelation function estimated from the results (**Sections 14.3.1** and **14.2.3**). The minimum lag (*i.e.*, time) between sampling events at which the autocorrelation is effectively zero can be used as an appropriate sampling interval.

## APPENDIX TABLES FOR PARAMETRIC RETESTING PLANS

The Unified Guidance provides tables of $\kappa$-multipliers for both interwell and intrawell prediction limits with retesting. It also provides separate tables for predicting individual future values versus future means. Four distinct retesting schemes are presented in the case of prediction limits for individual values: 1-of-2, 1-of-3, 1-of-4, and the modified California plan schemes. Five distinct schemes are presented for the case of future means: 1-of-1, 1-of-2, and 1-of-3 for means of order 2, and 1-of-1 and 1-of-2 for means of order 3.

Both the **Appendix D** interwell retesting tables (**Tables 19-1** through **19-9**) and the intrawell retesting tables (**Tables 19-10** through **19-18**) are similarly structured. Separate sub-tables are provided for a range of possible monitoring constituents ($c$ = 1 to 40) and for each of the retesting schemes mentioned above. Each table is divided into three parallel sections, one section applicable to annual statistical evaluations, one to semi-annual evaluations, and one to quarterly evaluations. Within each section, $\kappa$-multipliers are listed for all combinations of background sample size (from $n$ = 4 to 150) and number of wells (from $w$ = 1 to 200). These $\kappa$-multipliers are computed to meet a target annual SWFPR of 10%, as discussed in **Chapter 6**.

The **Appendix** tables also list those $\kappa$-multipliers which achieve adequate effective power compared to the ERPCs. The $\kappa$-multipliers are **bolded** when the effective power consistently exceeds the appropriate ERPC for mean level increases above background of 3 or more standard deviations

(designated as 'good' power). The multipliers are *italicized and shaded* when the effective power is somewhat less, but still consistently exceeds the ERPC at mean level increases of 4 or more standard deviations above background (designated as 'acceptable' power). Non-bolded, non-italicized entries achieve the target SWFPR, but have *low power*.

To use the tables, certain key statistical parameters should be known or identified. These include whether the prediction limit tests are interwell or intrawell, the evaluation schedule (annual, semi-annual, or quarterly), the number of constituents ($c$), the size of the background sample ($n$), and the number of compliance wells to be tested ($w$). In the interwell case, it is presumed that there are $n$ (upgradient) background measurements for each constituent ($c$). The listed $\kappa$-multiplier would then be applied to each of $c$ prediction limits, one for each monitoring constituent. The intrawell case presumes that there are $n$ well-specific background measurements designated at each well-constituent pair, thus giving $w \times c$ separate sets of intrawell background. Here, the κ-multiplier would be applied to each of $w \times c$ distinct prediction limits.

In situations where a mixture of test types is needed (*e.g.*, intrawell testing for some constituents, interwell for others), the Unified Guidance tables can still be employed. The $\kappa$-multipliers are computed to apportion an *equal share* of the overall cumulative SWFPR to each of the $w \times c$ tests that need to be run during a given statistical evaluation. Because of this fact, if $r$ of the constituents are analyzed using interwell tests, but $(c - r)$ of the constituents are handled using intrawell limits, correct prediction limits can be developed by first selecting an interwell κ-multiplier based on all $c$ constituents, and then selecting an intrawell $\kappa$-multiplier *also* based on $c$ constituents. This will ensure that the target SWFPR is met, although each multiplier is respectively applied only to a subset of the monitoring list.

Some background samples might be of different sizes, either for different constituents or at distinct wells (*e.g.*, when using intrawell background). Again the Unified Guidance tables can be inspected to select a different $\kappa$-multiplier for each distinct $n$. However, each multiplier should be chosen as if the background sample sizes were equal for all $w \times c$ tests. Thus, while a multiplier based on $n_1$ background observations is applied only to those tests involving that sample size, it should be selected from the **Appendix D** tables as if it will be applied to *all* the tests.

For network configurations not listed in **Tables 19-1** to **19-18** in **Appendix D**, an appropriate $\kappa$-multiplier can be estimated using bilinear interpolation. Such interpolation will be fairly accurate as long as adjacent table entries are used, representing the closest values to the desired combination of number of wells ($w$) and background sample size ($n$).

In general, to calculate a $\kappa_{w^*, n^*}$, where $w^*$ and $n^*$ are the desired input points that lie between the closest table entries as: $w_1 < w^* < w_2$ and $n_1 < n^* < n_2$, first calculate the fractional terms:

$$f_w = \frac{(w^* - w_1)}{(w_2 - w_1)} \quad \text{and} \quad f_n = \frac{(n^* - n_1)}{(n_2 - n_1)}$$

The interpolated κ-multiplier can then be computed as:

$$\kappa_{w^*,n^*} = (1 - f_w)(1 - f_n) \cdot \kappa_{w_1,n_1} + f_w(1 - f_n) \cdot \kappa_{w_2,n_1} + (1 - f_w) \cdot f_n \cdot \kappa_{w_1,n_2} + f_w \cdot f_n \cdot \kappa_{w_2,n_2} \qquad [19.12]$$

For example, suppose a $\kappa$-multiplier is needed for a 1-of-3 interwell prediction limit test for individual values using an annual evaluation schedule. Assume the monitoring network consists of $c = 5$ constituents monitored at $w = 28$ compliance wells, using $n = 17$ upgradient background measurements on which to base the prediction limit. From **Table 19-2** in **Appendix D**, the closest table entries, $\kappa_{,w,n}$ to the desired combination are $\kappa_{20,16} = 1.59$, $\kappa_{30,16} = 1.70$, $\kappa_{20,20} = 1.52$, and $\kappa_{30,20} = 1.62$. The interpolated value, $\kappa_{25,18}$, can then be found using the equations in [19.12]:

$$f_w = \frac{(28-20)}{(30-20)} = .8 \qquad f_n = \frac{(17-16)}{(20-16)} = .25$$

$$\kappa_{25,18} = (1-.8)(1-.25)\cdot\kappa_{20,16} + .8(1-.25)\cdot\kappa_{30,16} + (1-.8)\cdot.25\cdot\kappa_{20,20} + .8\cdot.25\cdot\kappa_{30,20}$$

$$= .15\cdot1.59 + .60\cdot1.70 + .05\cdot1.52 + .20\cdot1.62 \quad = 1.659$$

Important considerations in designing a reasonable retesting scheme for detection monitoring are discussed in **Chapter 6**. Given a background sample and a particular network configuration and size, parametric 1-of-$m$ plans tend to increase in statistical power as the order of $m$ increases. All of the schemes have greater power with larger background sample sizes ($n$). Furthermore, plans involving prediction limits for future means tend to be more powerful than similar plans using prediction limits for individual observations. So if the $\kappa$-multiplier for a particular plan is not **bolded** or *italicized*, another plan can be sought to achieve sufficient effective power using more resamples or perhaps changing to a mean prediction limit. Alternatively, the background sample size might need to be augmented if feasible, prior to implementing the retesting procedure.

### 19.3.1 TESTING INDIVIDUAL FUTURE VALUES

The advantages to using a prediction limit for future individual values include: 1) the ability to explicitly control the SWFPR across a series of well-constituent pairs; and 2) greater flexibility than that provided by prediction limits for future means (**Section 19.3.2**) to handle temporal autocorrelation. In those cases when the sampling frequency needs to be reduced to maximize statistical independence of the observations, the method can be applied to evaluations of a single new measurement (plus possible resamples) at each compliance point well.

To properly implement a prediction limit strategy for future values with retesting, it needs to be feasible to collect 2 to 4 independent measurements at each compliance well during a given evaluation period. All initial and any resamples are assumed to be statistically independent and thus should exhibit no autocorrelation.

If statistical evaluations are done annually, it may be possible to collect data on a quarterly basis and meet the minimal sampling requirements of any of the resampling schemes discussed in the Unified Guidance. However, more frequent evaluations (say semi-annual or quarterly) will require that new samples be collected perhaps monthly or every six weeks. In these cases, explicit tests for autocorrelation may need to be conducted before adopting a 1-of-$m$ retesting scheme with $m > 2$ or a

modified California plan. If significant autocorrelation is identified, the sampling frequency may need to be reduced and/or an alternate strategy utilizing fewer resamples may need to be adopted instead.

## PROCEDURE

Step 1. Identify the overall targeted annual false positive rate (SWFPR = $\alpha$ = 0.10). Determine the number of wells ($w$) to be monitored and the number of constituents ($c$) to be sampled at each well. Also determine whether the evaluation schedule at the unit or facility is *annual, semi-annual or quarterly*.

Step 2. Decide on the number of observations ($m$) to be predicted. To incorporate retesting, a maximum of two independent measurements should be collected from every compliance well during each evaluation period to use a 1-of-2 retesting scheme, three independent measurements if a 1-of-3 plan is desired, and four independent measurements if either a 1-of-4 plan or a modified California plan is employed.

Step 3. For interwell prediction limits given a background sample of $n$ measurements, compute the background sample mean ($\bar{x}$) and standard deviation ($s$) for each constituent. Then, based on the evaluation schedule (annual, semi-annual or quarterly), $c$, $n$, $w$, and the specific retesting scheme chosen, use **Tables 19-1** to **19-4** in **Appendix D** to determine a $\kappa$-multiplier possessing acceptable statistical power. Interpolate within the tables to find the closest multiplier if an exact value is not available.

For intrawell prediction limits, designate $n$ early measurements as intrawell background for each well-constituent pair; compute the intrawell background mean ($\bar{x}$) and standard deviation ($s$) for each case. Given the evaluation schedule, $c$, $n$, $w$, and the chosen retesting scheme, use **Tables 19-10** to **19-13** in **Appendix D** to determine an acceptably powerful $\kappa$-multiplier. Note: if the intrawell background sample size varies by well, a series of $\kappa$-multipliers should be computed, one for each distinct $n$.

For each $\kappa$-multiplier, calculate the upper prediction limit with (1– $\alpha$) confidence as:

$$PL_{1-\alpha} = \bar{x} + \kappa s \qquad [19.13]$$

If data were transformed prior to constructing the prediction interval, *back-transform* the prediction limit *before* making comparisons against the compliance point data. Unlike a prediction limit for future means, the formula for predicting $m$ future values *does not involve any transformation bias* if the comparison is made in the original measurement domain.

Step 4. Collect an initial measurement from each well-constituent pair being tested. Compare each value against either 1) the upper prediction limit based on upgradient background in the interwell case or 2) the intrawell prediction limit specific to that well-constituent pair. Depending on the retesting scheme chosen, if any initial compliance point concentration exceeds the limit, collect 1 to 3 additional resamples at that well. If feasible, analyze only for those constituents which exhibited initial exceedances. Compare these values sequentially against the upper prediction limit. If the test 'passes' prior to collection of all the scheduled resamples, the remaining resamples do not need to be gathered or compared against *PL*.

Step 5. Decide that the test at a given well passes (*i.e.*, the well is in-compliance) if any one or more of the resamples does not exceed *PL* when using a 1-of-*m* scheme or when at least 2 resamples do not exceed PL when using the modified California scheme. Identify the well as failing when either (1) *all* resamples using a 1-of-*m* plan also exceed the prediction limit, or (2) at least two of three resamples using a modified California plan exceed *PL*.

### ►EXAMPLE 19-1

A large hazardous waste facility with 50 compliance wells is to monitor 10 naturally-occurring inorganic parameters in addition to 30 non-naturally occurring volatile organic compounds that have never been detected on-site. Groundwater evaluations are performed on a semi-annual basis. If the regulating authority will allow up to two resamples per exceedence of the background concentration limit, construct an interwell prediction limit with adequate statistical power and false positive rate control on the following pooled set ($n$ = 25) of background sulfate measurements.

| BG Well | Sampling Date | Sulfate (mg/l) | Log (Sulfate) log(mg/l) |
|---|---|---|---|
| GW-01 | 07-08-99 | 63 | 4.143 |
| | 09-12-99 | 51 | 3.932 |
| | 10-16-99 | 60 | 4.094 |
| | 11-02-99 | 86 | 4.454 |
| GW-04 | 07-09-99 | 104 | 4.644 |
| | 09-14-99 | 102 | 4.625 |
| | 10-12-99 | 84 | 4.431 |
| | 11-15-99 | 72 | 4.277 |
| GW-08 | 10-12-97 | 31 | 3.434 |
| | 11-16-97 | 84 | 4.431 |
| | 01-28-98 | 65 | 4.174 |
| | 04-20-99 | 41 | 3.714 |
| | 06-04-02 | 51.8 | 3.947 |
| | 09-16-02 | 57.5 | 4.052 |
| | 12-02-02 | 66.8 | 4.202 |
| | 03-24-03 | 87.1 | 4.467 |
| GW-09 | 10-16-97 | 59 | 4.078 |
| | 01-28-98 | 85 | 4.443 |
| | 04-12-98 | 75 | 4.317 |
| | 07-12-98 | 99 | 4.595 |
| | 01-30-00 | 75.8 | 4.328 |
| | 04-24-00 | 82.5 | 4.413 |
| | 10-24-00 | 85.5 | 4.449 |
| | 12-01-02 | 188 | 5.236 |
| | 03-24-03 | 150 | 5.011 |

#### SOLUTION

Step 1. Assume for purposes of the example that there are no significant spatial differences among the well locations, either upgradient or downgradient. A check of normality of the pooled background sulfate measurements indicates that the interwell prediction limit should be constructed on the logged sulfate measurements rather than the raw concentrations.

Step 2. Groundwater evaluations must be conducted semi-annually (S). By excluding never-detected organic chemicals from the SWFPR calculation, the number of constituents that are to be considered is $c = 10$ at each of $w = 50$ wells.

Step 3. Since a maximum of two resamples will be allowed during any given evaluation period, neither the 1-of-4 nor the modified California retesting plan are an option. Consequently, only a 1-of-2 or 1-of-3 retesting strategy is appropriate. With $n = 25$ background measurements, **Tables 19-1** and **19-2** in **Appendix D** should be examined for a semi-annual evaluation schedule to determine $\kappa$-multipliers with adequate power. The multiplier of $\kappa = 2.75$ for a 1-of-2 plan has 'acceptable' power compared to the semi-annual ERPC, but the multiplier of $\kappa = 2.00$ for a 1-of-3 plan has 'good' power. Use the latter value to construct the interwell prediction limit.

Step 4. The sample log-mean and log-standard deviation of the sulfate background measurements are $\bar{y} = 4.32$ and $s_y = 0.376$, respectively. Use these values and the κ-multiplier to compute the prediction limit on the log-scale as

$$PL = \bar{y} + \kappa s_y = 4.32 + 2.00 \times 0.376 = 5.072$$

Then exponentiate the limit to back-transform it to the original measurement domain, for a final sulfate prediction limit of $PL = e^{5.072} = 159.5$ mg/l.

Step 5. Compare the final prediction limit against one new sulfate measurement from each of the 50 compliance point wells. For any exceedence, compare the first of two resamples to the prediction limit. If the limit is still exceeded, test the second resample. If all three measurements (initial plus two resamples) are above the prediction limit at any specific well, declare that a statistically significant exceedence for sulfate has been identified. If, however, neither of the resamples exceeds the limit, judge the evidence to be insufficient to declare the well to be out-of-compliance. ◄

►EXAMPLE 19-2

Due to significant natural spatial variability, an intrawell testing scheme needs to be adopted at a solid waste landfill that monitors for 5 inorganic constituents at each of 10 compliance wells. If only one year's worth of quarterly sampling data is available at each well, but no recent contamination is suspected, develop an appropriate modified California intrawell retesting plan for the following chloride measurements. Assume that one statistical evaluation must be conducted each year.

| Well ID | Chloride (mg/l) | Well Mean ± SD (mg/l) | Well ID | Chloride (mg/l) | Well Mean ± SD (mg/l) |
|---|---|---|---|---|---|
| GW-09 | 22 | 28.5 ± 10.021 | GW-16 | 31 | 43.6 ± 13.392 |
| | 18.4 | | | 34.6 | |
| | 39.9 | | | 60.1 | |
| | 33.7 | | | 48.7 | |
| GW-12 | 78 | 68.7 ± 7.208 | GW-24 | 23.4 | 33.98 ± 9.083 |
| | 70 | | | 36.4 | |
| | 61 | | | 31.1 | |
| | 65.8 | | | 45 | |
| GW-13 | 75.1 | 65.75 ± 8.128 | GW-25 | 33.5 | 31.38 ± 6.533 |
| | 65.6 | | | 30.2 | |
| | 67 | | | 23.1 | |
| | 55.3 | | | 38.7 | |
| GW-14 | 59.2 | 51.28 ± 8.427 | GW-26 | 79.8 | 60.92 ± 14.447 |
| | 57.1 | | | 61.3 | |
| | 41.1 | | | 57.8 | |
| | 47.7 | | | 44.8 | |
| GW-15 | 35 | 50.72 ± 15.672 | GW-28 | 37.7 | 38.0 ± 8.273 |
| | 56.8 | | | 26.6 | |
| | 69.8 | | | 45.7 | |
| | 41.3 | | | 42 | |

SOLUTION

Step 1. With $c = 5$ constituents, $w = 10$ wells, one annual evaluation, and an intrawell background size for each well of only $n = 4$, **Table 19-13** in **Appendix D** can be examined to locate a possible $\kappa$-multiplier, leading to an interpolated $\kappa = 4.33$. Although this multiplier will adequately control the annual SWFPR to 10% or less, it yields low power for identifying contamination. As an alternative, try computing a pooled standard deviation across the compliance wells for chloride.

Step 2. Side-by-side box plots (**Section 11.1**) of the chloride values exhibit no obvious differences in spread or variation. The *F*-statistic for Levene's test (**Section 11.2**) is also non-significant ($F = 1.0673$) at the $\alpha = 5\%$ level, suggesting that the variances are not unequal and that a pooled standard deviation can be appropriately formed.

Step 3. Conduct a one-way ANOVA on all chloride measurements from the 10 compliance wells, using Wells as the main factor (**Section 13.2.2**). The ANOVA table is presented below.

| Source of Variation | Sums of Squares | Degrees of Freedom | Mean Squares | *F*-Statistic |
|---|---|---|---|---|
| Between Wells | 7585.25 | 9 | 842.81 | 7.55 |
| Error (within wells) | 3350.37 | 30 | 111.68 | |
| Total | 10935.62 | 39 | | |

Step 4. Compute the square root of the Error Mean Squares (also called the *root mean squared error* or RMSE) component in the ANOVA table to derive an estimate of the pooled intrawell standard deviation of $s_p$ = 10.568. This estimate of the average intrawell variation has 30 degrees of freedom [*df*], computed by multiplying (4–1) = 3 degrees of freedom per well times the number of wells, or $df = 3 \times 10 = 30$.

Step 5. The **Appendix D** tables are not used to derive $\kappa$-multipliers when a pooled standard deviation estimate is used for intrawell prediction limits. **R** script listed in **Appendix C** is used (see **Section 13.3**). For a modified California retesting strategy with n = 4 and *df* = 30, the κ-multiplier becomes $\kappa$ = 1.98.[7] This value not only controls the SWFPR but also has good statistical power. So use this multiplier along with the pooled intrawell standard deviation to compute an intrawell prediction limit for each compliance well. As an example, since the mean for chloride at well GW-09 is 28.5, the intrawell prediction limit would be:

$$PL = 28.5 + 1.98 \times 10.568 = 49.4 \text{ mg/l}$$

Prediction limits for the other compliance wells would be computed similarly. ◄

## 19.3.2 TESTING FUTURE MEANS

### BACKGROUND AND REQUIREMENTS

The background, requirements, and assumptions for a prediction limit on future means of order *p* are essentially identical to those for prediction limits for future values (**Section 19.3**). For a comparable level of sampling effort, predicting a future mean offers *increased effective power* compared to a strategy that uses prediction limits for individual future values. To properly implement a prediction limit strategy for future means with retesting, *it must be feasible to collect 2 to 6 independent measurements at each compliance well during a given evaluation period*. All initial and resample measurements are assumed to be statistically independent.

To include explicit retesting, it should be feasible to collect either 2*p* or 3*p* independent measurements per well during each evaluation. The initial *p* observations are used to form the initial mean, while the remaining values are used to form either one or two resample means. If statistical evaluations are done annually, it may be possible to collect quarterly data and meet the minimal sampling requirements for *p* = 2 and a 1-of-2 retesting scheme. For more frequent semi-annual or quarterly evaluations, a larger order *p* or a retesting scheme entailing two resample means will require that new samples be collected perhaps monthly or every six weeks. An explicit test for autocorrelation should be made before adopting the strategy presented here. If significant autocorrelation exists, the frequency of sampling may need to be reduced and alternate prediction limit strategies considered such as a 1-of-1 plan for a future mean (see **Section 19.1**) or individual future values (**Section 19.3.1**).

[7] The EPA Region 8 approximation equation described in **Chapter 13, Section 13.3** provides a κ-multiple estimate of 1.99 for individual wells at *n* = 4. The annual κ-factor for *w* = 10 and *c* =5 and *n* = 31 in **Table 19-13** of **Appendix D** is interpolated as $\kappa$ = 1.508. Using the appropriate *A, b* & *c* coefficients from Chapter 13, Note 2 for the modified California plan, results are quite close to that generated from **R**-script.

An important difference between testing means versus individual values is that in some cases it may not be necessary to implement a retest at all. As noted above, for the same degree of sampling effort, a prediction limit for a mean of two or more observations can provide greater effective power than a prediction limit for the same number of individual values, *even if a resampled mean is not collected*. In other words, when a 1-of-2 retesting plan for individual observations is compared to a 1-of-1 plan for means of order 2, the 1-of-1 mean-based scheme generally has greater power for identifying real concentration increases if background samples sizes are $n > 10$ (compare $\kappa$-multiple power ratings at higher $n$, $c$, and $w$ in **Tables 19-1** and **19-5 *of* Appendix D**) A similar comparison holds between a 1-of-3 retesting plan for individual observations and a 1-of-1 plan for a mean of order 3 (**Table 19-2** versus **Table 19-8** in **Appendix D**).

Even more powerful prediction limits for future means are possible when explicit retesting is added to the procedure. However, the minimum sampling increases substantially. With a 1-of-2 retesting plan for means of order 2, as many as four independent groundwater measurements needs to be collected and analyzed per evaluation period. With a 1-of-3 plan for means of order 2 or a 1-of-2 plan for means of order 3, the sampling increases to as many as six independent observations per period. The latter plans may only be feasible for a single annual evaluation.

A problem common to all future mean prediction limits arises if the data have to be normalized via a transformation. In this case, all comparisons need to be made on the transformed data in order to avoid a transformation bias. As a consequence, the procedure is not a direct test of the background and compliance point *arithmetic* means. The test is still valid as a measure of significant mean differences in the transformed domain (*e.g.*, a test of geometric mean differences for logarithmic data). To the extent that the populations being compared share a common variance in the transformed domain, it may also indicate that a significant difference on the transformed scale also corresponds to a significant difference in the arithmetic means of the original populations.

A final potential drawback is that although a 1-of-$m$ plan for future observations and a 1-of-1 plan for means of order $p = m$ seem to require the same total sampling effort, a prediction limit for observations can actually entail *less* sampling. For a future mean test of order $p = m$, $m$ individual measurements will always need to be collected and analyzed. With a prediction limit for individual observations, the first sample is analyzed and compared to the limit. If it passes (*i.e.*, does not exceed the limit) there is no need to test the second or subsequent observations. Any subsequent resample that passes, also indicates that no further resample comparisons are needed for that test.

Under typical conditions at a site where most or all tested well-constituent pairs are likely to be at background conditions, there is a substantial savings in the number of samples for future observations versus means of the same size. It can also be noted that the same principle is true for a 1-of-2 test of a mean of order 2. Under background conditions, the two initial mean samples may be all that is required. When groundwater is contaminated, both the 1-of-$m$ retesting plan for observations and the 1-of-1 plan for a mean of order $p = m$ require exactly the same amount of sampling and analysis to identify a significant exceedance.

## PROCEDURE

Step 1. Identify the number of wells ($w$) to be monitored and the number of constituents ($c$) to be sampled at each well. Also identify the evaluation schedule as annual (A), semi-annual (S), or quarterly (Q).

Step 2. Decide on the order ($p$) of the future mean to be predicted. To incorporate retesting, it needs to be possible to collect $2p$ independent samples during each evaluation period to use a 1-of-2 retesting scheme, or $3p$ independent samples if a 1-of-3 plan is desired.

Step 3. If an *interwell* prediction limit is needed, use the common sample of $n$ (upgradient) background measurements to compute the background sample mean ($\bar{x}$) and standard deviation ($s$). Given the $n$ background measurements, $w$, $c$, $p$, and the evaluation schedule (annual, semi-annual or quarterly), use **Tables 19-5 to 19-9** in **Appendix D** to determine a $\kappa$-multiplier possessing acceptable statistical power. Calculate the upper prediction limit on background as:

$$PL = \bar{x} + \kappa s \qquad [19.14]$$

If *intrawell* prediction limits are needed, designate $n$ early measurements at each compliance well as intrawell background. Compute the background sample mean ($\bar{x}$) and standard deviation ($s$) for each well. Then, based on $n$, $w$, $c$, $p$, and the number of evaluations per year, use **Tables 19-14** to **19-18** in **Appendix D** to determine an adequately powerful $\kappa$-multiplier. Compute an intrawell prediction limit for each compliance well using equation [19.14]. Note: if the intrawell background sample sizes vary by well, a series of $\kappa$-multipliers will need to be identified in these **Appendix D** tables, one for each distinct $n$.

If the background data were transformed prior to constructing the prediction limit, *also transform any compliance point data before making comparisons against the prediction limit*. In particular, compute the comparison mean of order $p$ using the *transformed values*, rather than transforming the sample mean of the raw concentrations.

Step 6. Collect $p$ initial measurements from each compliance well. Compute the mean of order $p$ for each well, *first transforming the data if necessary using the same function applied to background*. Then compare each mean against the upper prediction limit. If retesting is desired, for any compliance point mean that exceeds the limit, collect either $p$ or $2p$ additional resamples at that well, depending on the retesting scheme chosen. Form either one or two resample means of order $p$ from these data; compare these means sequentially to the upper prediction limit.

Step 7. Identify the well as potentially contaminated when either 1) the initial mean of order $p$ exceeds the limit in a 1-of-1 plan, or 2) the initial mean and *all* resample means using a 1-of-2 or 1-of-3 plan also exceed the prediction limit. Deem the well to be in-compliance if either 1) the initial mean does not exceed the prediction limit, or 2) *any* of the resample means do not exceed the limit.

## ►EXAMPLE 19-3

Suppose a large facility with minimal natural spatial variation is to monitor for 20 separate naturally-occurring inorganic chemicals along with a number of other never detected organic constituents. If 100 compliance wells are to be tested every six months and 25 background sample measurements are available, which resampling plans can control the SWFPR, providing acceptable statistical power? Assume that the data for each inorganic compound can be normalized and that the temporal autocorrelation between successive samples at the same well is minimal, *provided* that no more than four samples are collected during any semi-annual period.

### SOLUTION

Step 1. The frequency of statistical evaluations is semi-annual (S). Excluding never-detected compounds from the SWFPR calculation leaves $c$ = 20 constituents that need to be explicitly tested at each of $w$ = 100 wells. For each of these constituents, since the data can be normalized, assume that an interwell prediction limit can be constructed using $n$ = 25 background measurements.

Step 2. Determine $\kappa$-multipliers and power ratings for seven possible prediction limit retesting plans excluding the 1-of-3 mean order 2 and the 1-of-2 mean order 3 tests. Use the sub-tables identified as "*20 COCs, Semi-Annual*" for $n$ =25 and $w$ = 100 in interwell **Tables 19-1** through **10-9** in **Appendix D**, to obtain the following:

| Prediction Limit Plan | $\kappa$-Multiplier | Power | Total Samples |
|---|---|---|---|
| 1-of-2, observations | 3.13 | Low | 2 |
| 1-of-3, observations | 2.31 | Good | 3 |
| 1-of-4, observations | 1.81 | Good | 4 |
| Mod. California, observations | 2.54 | Good | 4 |
| 1-of-1, mean order 2 | 3.56 | Acceptable | 2 |
| 1-of-2, mean order 2 | 2.29 | Good | 4 |
| 1-of-1, mean order 3 | 2.95 | Good | 3 |

Step 3. Compare the various plans in terms of statistical power and typical sampling effort. The only plan with low power is the 1-of-2 scheme for observations. The 1-of-1 mean order 2 has acceptable power. The other plans all have good power (*i.e.*, ones consistently meeting or bettering the ERPC for mean-level increases above background of 3 or more standard deviations), but potentially require either 2 or 3 resamples.

Restricting attention to those with good power, the least potential sampling effort is required by the 1-of-1 plan for a mean of order 3 or a 1-of-3 plan for observations. These two plans would requires *less total* sampling than the 1-of-4 plan for observations, the 1-of-2 mean order 2 plan and the *same or less* sampling than the modified California plan for observations in identifying a contaminant release.

If groundwater is not contaminated, the 1-of-$m$ plans for observations require a minimum of 1 measurement to demonstrate that the well is in-bounds (*i.e.*, when the initial measurement does not exceed the background limit) as does the modified California plan. The 1-of-2 plan for a mean of order 2 requires a minimum of 2 measurements, and the 1-of-1 plan for a mean of order 3 requires a minimum of 3 measurements. On balance, the 1-of-3 plan for individual observations or the 1-of-2 plan for a mean of order 2 may provide the best compromise

between minimizing sampling effort and offering a higher probability of identifying contaminated groundwater. ◄

►EXAMPLE 19-4

Use the chloride data of **Example 19-2** to compute and contrast prediction limits for a future mean of order 2, with and without explicit retesting. Assume as before that 10 wells are monitored for 5 inorganic constituents, and evaluated on an annual basis.

SOLUTION

Step 1. The chloride data in **Example 19-2** showed significant spatial variability, suggesting the use of intrawell prediction limits. Furthermore, a one-way ANOVA evaluation of the $w$ = 10 compliance wells indicated that a pooled standard deviation estimate of $s_p$ = 10.568 with 30 degrees of freedom could be used to build intrawell prediction limits, instead of using individual variance estimates from each compliance well.

Step 2. With $c$ = 5 constituents, $w$ = 10 wells to be monitored, one annual evaluation (A), and a pooled degrees of freedom of $df$ = 30, the **R** script in **Appendix C** can be repeatedly run to determine $\kappa$-multipliers for each retesting scheme for prediction limits on means of order 2. Since the sample size for each of the 10 wells is the same $n$ = 4, the following multiples were generated from the **R**-script for the 1-of-1 to 1-of-3 tests of mean order 2: $\kappa$ = 2.68, 1.88 and 1.51, respectively.[8] The prediction limits can then be constructed using equation [19.15], as shown for the first five compliance wells in the table below.

$$PL = \bar{x} + \kappa s_p \qquad [19.15]$$

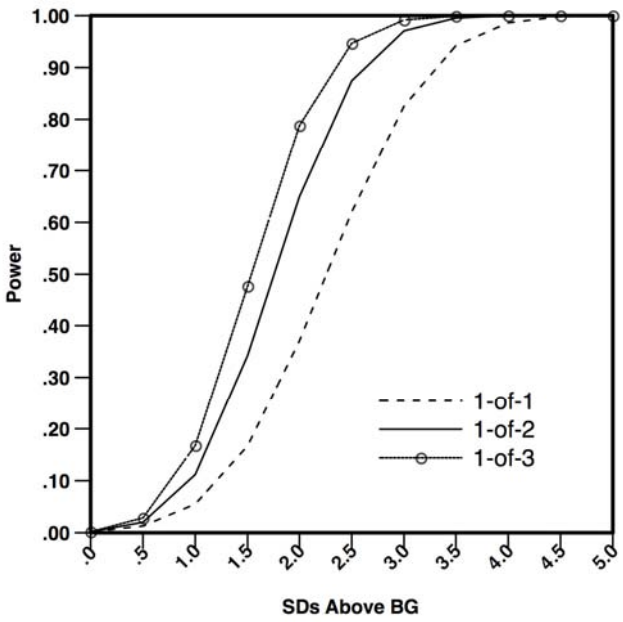
Step 3. While the power of each retesting plan is rated 'good' compared to the annual-evaluation ERPC, the prediction limits are obviously higher when less (or no) explicit retesting is conducted. Depending on conditions at the site, the range of approximately 13 mg/l of chloride in the well-specific prediction limits may or may not be important in deciding which strategy to use. The 1-of-1 plan for a mean of order 2 requires fewer total samples than the other plans. In some situations, the higher initial limits may be outweighed by the savings in sampling cost.

On the other hand, the ERPC provides a minimal standard for assessing statistical power. There can be a range of power curves even among plans all rated as 'good' seen in **Figure 19-1** below, where the full effective power curves for these three strategies are presented. Clearly, the 1-of-2 and 1-of-3 plans for means of order 2 have visibly higher power than the 1-of-1 retesting scheme. If site conditions permit, it may be beneficial to incorporate the 1-of-2 plan as a reasonable compromise between the gain in statistical power versus the increase in sampling requirements (for contaminated wells). ◄

[8] Using the Region 8 approximation equation in **Chapter 13**, the corresponding $\kappa$-multiples were 2.69, 1.89 and 1.52, respectively, based on tabular values at $n$ = 31 of 2.258, 1.364 & .946 and using the appropriate $A$, $b$ & $c$ coefficients for each test. Results are very comparable to the R-script values.

| Well ID | Retesting Plan | κ Multiplier | Power Rating | Well Mean (mg/l) | Prediction Limit |
|---|---|---|---|---|---|
| GW-09 | 1-of-1 | 2.68 | Good | 28.50 | 56.82 |
| | 1-of-2 | 1.88 | Good | 28.50 | 48.37 |
| | 1-of-3 | 1.51 | Good | 28.50 | 44.46 |
| GW-12 | 1-of-1 | 2.68 | Good | 68.70 | 97.02 |
| | 1-of-2 | 1.88 | Good | 68.70 | 88.57 |
| | 1-of-3 | 1.51 | Good | 68.70 | 84.66 |
| GW-13 | 1-of-1 | 2.68 | Good | 65.75 | 94.07 |
| | 1-of-2 | 1.88 | Good | 65.75 | 85.62 |
| | 1-of-3 | 1.51 | Good | 65.75 | 81.71 |
| GW-14 | 1-of-1 | 2.68 | Good | 51.28 | 79.60 |
| | 1-of-2 | 1.88 | Good | 51.28 | 71.15 |
| | 1-of-3 | 1.51 | Good | 51.28 | 67.24 |
| GW-15 | 1-of-1 | 2.68 | Good | 50.72 | 79.04 |
| | 1-of-2 | 1.88 | Good | 50.72 | 70.59 |
| | 1-of-3 | 1.51 | Good | 50.72 | 66.68 |

Figure 19-1. Comparison of Power Curves for 1-of-m Plans for Mean of Order 2



# 19.4 NON-PARAMETRIC PREDICTION LIMITS WITH RETESTING

## BACKGROUND AND PURPOSE

When parametric prediction limits are not appropriate, either due to a large fraction of non-detects or data that cannot be normalized, retesting can be conducted using *non-parametric* prediction limits. The Unified Guidance discusses retesting schemes for both individual future values and for future medians (in parallel to the parametric options discussed in **Section 19.3**). Tests on individual observations include the three 1-of-*m* plans and modified California plan approaches. Tests on future medians include the 1-of-1 and 1-of-2 plans for medians of order 3. The basic strategy is to establish a non-parametric prediction limit for each monitoring constituent based on background measurements so that it accounts for the number of well-constituent tests in the overall network. Instead of determining a $\kappa$-multiplier, a non-parametric limit is computed as an *order statistic* from the background sample. The term order statistic refers to one of the values in a sorted (or *ordered*) data set.

In order to maintain adequate statistical power while minimizing the overall false positive rate, retesting will almost always be needed as part of the detection monitoring system design. As in the parametric case, a specific number of additional, *independent* resamples will potentially need to be collected for each compliance well test. The initial and subsequent resamples are then compared against the non-parametric prediction limit.

The largest or second-largest value in background is often selected as a non-parametric limit, representing the $n$th or $(n-1)$th order statistics. With higher level 1-of-*m* tests of observations, an even lower order statistic may be more appropriate in achieving an optimal balance between the desired SWFPR and adequate statistical power. This can be particularly true if the background sample size is large, but depends on the overall network design requirements. Although the Unified Guidance provides tables of non-parametric limits only for the largest and second-largest order statistics, EPA Region 8 has released software written in Visual Basic® labeled the *Optimal Rank Values Calculator* that computes the optimal choice of order statistic for 1-of-*m* retesting plans for $m = 1$ to 4. The program also provides approximate statistical power estimates based on user inputs of a target cumulative false positive rate, background sample size, and number of simultaneous tests to be conducted. The software and explanatory narrative will be provided on the EPA website.[9]

## REQUIREMENTS AND ASSUMPTIONS

When more independent data are added to the testing procedure, retesting with non-parametric prediction limits leads to more powerful and more accurate assessments of possible contamination. As with parametric retesting schemes, a balance must be struck between 1) quick identification and confirmation of contaminated groundwater and 2) statistical independence of successive resamples. All retesting strategies depend on the assumption of statistical independence between successive resamples. This trade-off is typically resolved by allowing enough time between resamples to allow both the well to

[9] The calculator, an accompanying narrative, fact sheet and this guidance will be located on the EPA website: http://www.epa.gov/hazard/correctiveaction/resources/guidance/sitechar/gwstats/index.htm. If the calculator cannot be accessed, contact Mike Gansecki for assistance (e:mail: gansecki.mike@epa.gov; or phone: 303- 312-6150.)

recharge and additional groundwater to flow past the well screen, and by limiting the number of possible resamples to 2 or 3.

Non-parametric retesting schemes offer somewhat less flexibility than their parametric counterparts. As with other non-parametric statistical intervals, the same SWFPR control afforded by a parametric interval based on a small *n* cannot usually be attained in a non-parametric interval; larger sample sizes are almost always necessary. $\kappa$-multipliers for parametric prediction limits are continuous statistical parameters that can be adjusted to match a desired false positive rate for even the smallest sample sizes. By contrast, the bounds of non-parametric intervals are restricted to values in the observed background sample. For a given sample size and number of tests to be run, any order statistic selected from background as the non-parametric prediction limit results in a *discrete probability* of false positive error. Altering the prediction limit by selecting a different order statistic changes the false positive rate only in discrete probability steps, providing a less efficient means of controlling the SWFPR.

The non-parametric prediction limit tests provided in the Unified Guidance do not require the underlying distribution to be normal. One potentially attractive application is for background data sets containing higher percentages of non-detects which cannot be normalized. For some constituent data sets, it may be possible to pool data from several upgradient and historical compliance wells to generate much larger total background sizes. A non-parametric Kruskal-Wallis test of medians can establish that these data are appropriate for pooling.

Since larger background sample sizes are needed because no distributional model is posited, the non-parametric testing schemes are *most applicable to interwell* comparisons. Small *intrawell* background sample sizes make it difficult for any of the non-parametric test options to be applied which can meet the SWFPR cumulative false positive design objective. Unlike parametric intrawell tests, effective sample sizes cannot be expanded by estimating a common pooled standard deviation across a number of wells. This conclusion is generally true no matter what order statistic is used to estimate the non-parametric prediction limit. But there are other considerations which might allow intrawell testing using non-parametric alternatives. For a given sample size, target false positive, a fixed maximum and number of total tests, the higher 1-of-*m* tests of future observations will have lower achievable false positive errors, with the 1-of-4 test the lowest. If the background sample size is increased through periodic additions, this false positive will continue to drop. The power of these tests using the maximum with small sample sizes is almost always greater than the EPA reference levels. A temporary strategy might be to utilize the highest order 1-of-*m* test for intrawell purposes until larger sample sizes are available. However, the target cumulative false positive rate may not initially be met. With larger sample sizes, it may also be possible to decrease the *m* of the test and still achieve the target false positive rate.

Even interwell comparisons between upgradient and downgradient wells are acceptable only if the degree of spatial variability is insignificant. Fortunately, spatial variability may be less of a problem in those cases where a non-parametric retesting scheme might be implemented, *i.e.*, when the detection rate of the chemical being monitored is fairly low. High constituent non-detect rates tend to result in more uniform spatial distribution across site wells, allowing for similar median concentrations.

## APPENDIX TABLES FOR NON-PARAMETRIC PREDICTION LIMITS

To design appropriate non-parametric prediction limits with retesting, the Unified Guidance provides separate tables for predicting individual future values versus future medians. Four distinct retesting schemes are presented in the case of prediction limits for individual values: 1-of-2, 1-of-3, 1-of-4, and modified California plan schemes. Two distinct schemes are presented for the case of future medians: 1-of-1 and 1-of-2 for medians of order 3.

Unlike the tables for parametric prediction limits discussed in **Section 19.3**, non-parametric prediction limits do not involve $\kappa$-multipliers. Instead, the entries in **Tables 19-19** to **19-24** of **Appendix D** consist of *per-constituent significance levels*. These levels represent the *achievable* false positive rate ($\alpha_{const}$) associated with each tested constituent for a given retesting scheme, choice of non-parametric prediction limit, and network configuration (*i.e.*, number of wells [*w*] and background sample size [*n*]).[10] The non-parametric prediction limit can be estimated via any order statistic from the background sample. However, the most practical limits are usually either the maximum observed background value or the second-highest value. Consequently, the Unified Guidance provides tables for these two options.

Each table for the six specific non-parametric tests contains two sub-tables. One uses a limit based on the background maximum and the other the second-highest background value. All the tables are otherwise similarly structured. Within each table and sub-tables, per-constituent significance levels are given for all combinations of background sample size ($n$ = 4 to 200) and number of wells ($w$ = 1 to 200). These significance levels can be used to meet a target annual SWFPR of 10%, discussed in **Chapter 6**.

Correct use of these tables involves a few important considerations. First, if an *interwell* prediction limit is desired, the *target per-constituent* false positive rate ($\alpha_{const}$) needs to be computed. Any prediction limit strategy selected should have a table entry no greater than $\alpha_{const}$ in order to ensure that the annual SWFPR is no greater than 10%. To compute this target rate, use the formula:

$$\alpha_{const} = 1 - (1 - \alpha)^{1/c} \qquad [19.16]$$

where $c$ equals the number of monitoring constituents and $\alpha$ is the SWFPR = 0.10.

Unlike the tables for parametric prediction limits, separate tables are not provided for each of the three most common evaluation schedules (*i.e.*, annual, semi-annual, and quarterly). The number of 'wells' in each non-parametric table must be regarded as the actual number of compliance wells ($w$) times the number of annual statistical evaluations ($n_E$ = 1, 2, or 4). For using these tables, let $w^* = w \times n_E$. This adjustment is necessary because on each evaluation, $w$ wells should be compared against a prediction limit computed from a common interwell background. A site with $w^*$ wells tested annually is statistically equivalent to a site having $w$ distinct well locations tested $n_E$ times per year ($w \times n_E$ tests).

[10] Per-constituent rates instead of network-wide false positive rates are given in these tables and those of Davis and McNichols (1994; 1999) for computational reasons. Although the mathematical algorithm is exact, it is difficult to compute with accuracy for a large number of tests ($r$). Hence the decomposition of $r$ into constituents ($c$) times wells ($w$). By calculating the per-constituent false positive rate, only the number of wells ($w$) need be varied.

Once *w** is computed in this way, the table entry corresponding to *w** and *n* represents the *achievable* annual false positive rate per constituent. As noted, this rate should not exceed the target rate ($\alpha_{const}$) in order to meet the overall SWFPR. If $\alpha_{const}$ is exceeded for a given choice of retesting scheme and choice of non-parametric prediction limit, a different limit or scheme should be considered. In general, selecting a 1-of-*m* retesting scheme with larger *m* will lead to a lower achieved false positive rate. Also, per-constituent significance levels for the modified California approach are generally larger than those for the 1-of-*m* plans.

If *intrawell* prediction limits are needed, a somewhat different method needs to be employed to correctly use the per-constituent significance levels in **Tables 19-19** through **19-24** of **Appendix D**. In this case, a target *per well-constituent pair* false positive rate ($\alpha_{w\cdot c}$) needs to be first computed using the equation:

$$\alpha_{w\cdot c} = 1 - (1-\alpha)^{1/(w\cdot c)} \qquad [19.17]$$

where $\alpha$ is the SWFPR, *w* equals the actual number of compliance wells and *c* is the number of monitoring constituents. Then the placeholder *w** for the non-parametric tables is to be equated with the number of annual statistical evaluations (*w** = $n_E$ = 1, 2, or 4). *w** represents the number of times per year that the common intrawell background at any given well-constituent pair will be compared against new compliance measurements from that well. The table entry corresponding to *w** and the intrawell background sample size *n* may be regarded as the achievable false positive rate per well-constituent pair. This rate should not exceed the target rate, $\alpha_{w\text{-}c}$, if the overall SWFPR is to be met.

The same approach presented in **Section 19.3** is used if a mixture of test methods is needed (*e.g.*, parametric prediction limits for some constituents, and non-parametric limits for other constituents). By construction, the target SWFPR is evenly proportioned across the list of monitored constituents. As long as the significance level per constituent (interwell case) or per well-constituent pair (intrawell case) is computed using all *c* constituents and not just those for which a non-parametric prediction limit test will be applied, the SWFPR will not exceed $\alpha = 0.10$ on an annual basis.

**Tables 19-19** through **19-24** in **Appendix D** provide the same **bold**, *italicized* or plain text used to identify 'good', 'acceptable' and 'low' power ratings following the ERPC 3 and 4 standard deviation reference criteria as in the parametric prediction limit tables.

As final technical notes about these tables, the significance levels listed as table entries are presented using a short-hand notation in order to compactly present a wide range of false positive rates. In this notation, the first four non-zero digits of the significance level are given, followed if necessary, by the symbol *–d*. The value *d* represents the number of leading zeros to the right of the decimal point. This is equivalent to taking the non-zero portion of the entry and multiplying it by $10^{-d}$ to get the actual significance level. As an example, if the entry is .4251–4, the equivalent significance level is .00004251. Entries without the *–d* symbol are the actual fractional significance levels where no adjustment is needed.

For network configurations (number of wells [*w*] and background sample size [*n*]) not listed in **Tables 19-19** through **19-24** in **Appendix D**, bilinear interpolation can be used to approximate the significance level associated with the desired configuration. As discussed in **Section 19.3**, interpolation

should be restricted to the closest four adjacent table entries. The shorthand significance level notations in the tables should first be converted to actual fractions before interpolating.

## 19.4.1 TESTING INDIVIDUAL FUTURE VALUES

### BACKGROUND AND REQUIREMENTS

The Unified Guidance recommends two variations of non-parametric prediction limits for use in groundwater detection monitoring. The first is the prediction limit for individual future values, introduced in **Section 18.3.1**. The other is the prediction limit for future medians, detailed in **Section 18.3.2**. Basic requirements for non-parametric prediction limits are outlined in those sections.

The main advantage to a prediction limit for future values is its overall flexibility and ease of implementation. Fewer data from each compliance well are needed to implement the test compared to a prediction limit for a future median. Only an initial observation from each compliance point may be needed to identify a well-constituent pair 'in-bounds'; initial exceedances can be followed by up to a maximum of three additional individual resamples. Once the non-parametric upper prediction limit has been selected from background as a large order statistic (often the maximum or second-largest value), each compliance point measurement is compared directly against this upper limit.

The user should decide which retesting scheme to use and how many resamples per well are feasible, given that the measurements from any well during a given evaluation period need to be statistically independent. **Tables 19-19** through **19-22** in **Appendix D** can be employed to compare the *achievable false positive rates* of different schemes and to determine whether they exhibit adequate effective power. The user can also explore EPA Region VIII's *Optimal Rank Values Calculator* software to consider order statistics other than the maximum or second-largest.

### PROCEDURE

Step 1. For an *interwell* test, use the number of monitoring constituents (*c*) in equation [19.16] to determine the target per-constituent false positive rate ($\alpha_{const}$). Also multiply the number of yearly statistical evaluations ($n_E$) by the actual number of compliance wells (w) to determine the look-up table entry, *w**. Then depending on the background sample size *n* and *w*, choose a type of non-parametric prediction limit (*i.e.*, maximum or 2nd highest value in background) and a retesting scheme for individual observations using **Tables 19-19** through **19-22** in **Appendix D**. The final plan should have an achieved significance level no greater than $\alpha_{const}$ and also should be labeled with 'acceptable' or 'good' power in the **Appendix** tables.

Step 2. For an *intrawell* test, use the number of constituents (*c*) and the actual number of compliance wells (*w*) in equation [19.17] to compute the target significance level per well-constituent pair ($\alpha_{w\text{-}c}$). Set *w** in the look-up table equal to the number of yearly evaluations, $n_E$. Based on $w^* = n_E$ and the intrawell background sample size *n*, choose a non-parametric prediction limit and retesting scheme so that the achieved well-constituent pair significance level (*i.e.*, the selected table entry) does not exceed the target significance level, $\alpha_{w\text{-}c}$, and also is labeled with 'acceptable' or 'good' statistical power.

Step 3. Sort the background data into ascending order and set the upper prediction limit equal to an appropriate order statistic of the data (***e.g.***, the maximum or the second-largest observed value). If all constituent measurements in a background sample are non-detect, use the Double Quantification rule in **Chapter 6**. The constituent should not be included in calculations for identifying the target false positive.

Step 4. Collect one initial measurement per compliance well. Then compare each initial measurement against the upper prediction limit. Depending on the retesting scheme chosen, for any compliance point value that exceeds the limit, collect one to three additional resamples from that well. Again compare the resamples against the upper prediction limit.

Step 5. Identify any well with an initial exceedance as potentially contaminated when either (1) ***all*** resamples using a 1-of-2, 1-of-3, or 1-of-4 plan also exceed the prediction limit, or (2) at least two resamples exceed the limit using a modified California retesting scheme. Conversely, declare a well to have 'passed' the test if either 1) the initial measurement does not exceed the prediction limit, 2) ***any*** resamples from a 1-of-***m*** scheme do not exceed the limit, or 3) at least 2 of 3 resamples from a modified California approach do not exceed the limit.

## 19.4.2 TESTING FUTURE MEDIANS

### BACKGROUND AND REQUIREMENTS

Prediction limits for a future median based on either a single or with one repeat (1of-1 or 1-of-2 tests) are two non-parametric procedures recommended as retesting methods in the Unified Guidance. Compared to a prediction limit for future individual values, the prediction of a median (**Chapter 18**) often requires more data to be collected from each compliance well particularly if resampling is included. Slightly greater statistical manipulation is also needed once the data are in hand. For the 1-of-1 test, the initial median to be predicted requires at least two initial observations from each compliance point, and any resample medians will require additional sets of up to three measurements, all of which needs to be statistically independent.

Given equal amounts of data and the same input conditions, a prediction limit for a future median tends to be more statistically powerful than a prediction limit for individual values. This is true whether one uses a fixed order statistic or selects across a range of order statistics to form the prediction limit. Because of this and the fact that both spatial variability and autocorrelation may be less of a problem (or at least less easily assessed) when the detection rate is low and a non-parametric strategy is needed, the Unified Guidance includes **Appendix D** tables for both a 1-of-1 scheme and a 1-of-2 scheme to predict medians of order 3. The 1-of-2 median test will have a lower achievable false positive rate than the 1-of-1 version, with all other conditions equal.

Depending on the number of annual evaluations and the test configuration, care needs to be taken that potentially needed samples are far enough apart in time. The series of observations from any well is assumed to be uncorrelated. If autocorrelation is a problem, a prediction limit for future values (**Section 19.4.1**) should be considered in which the per-well sampling requirements with explicit retesting are more modest.

## PROCEDURE

Step 1. For an *interwell* test, use the number of monitoring constituents (*c*) in equation [19.16] to determine the target per-constituent false positive rate ($\alpha_{const}$). Also multiply the number of yearly statistical evaluations ($n_E$) by the actual number of compliance wells (w) to determine the look-up table margin value, *w**. Then, depending on the background sample size *n* and *w**, choose a type of non-parametric prediction limit (*i.e.*, maximum or 2nd highest value in background) and a retesting scheme for future medians using **Tables 19-23** to **19-24** in **Appendix D**. The final plan should have an achieved significance level no greater than $\alpha_{const}$, and also should be labeled with 'acceptable' or 'good' power in the **Appendix** tables.

Step 2. For an *intrawell* test, use the number of constituents (*c*) and the actual number of compliance wells (*w*) in equation [19.17] to compute the target significance level per well-constituent pair ($\alpha_{w\text{-}c}$). Set *w** in the look-up table margin equal to the number of yearly evaluations, $n_E$. Based on $w^* = n_E$ and the intrawell background sample size (*n*), choose a non-parametric prediction limit and retesting scheme for future medians so that the achieved well-constituent pair significance level (*i.e.*, the selected table entry) does not exceed the target significance level, $\alpha_{w\text{-}c}$, and also is labeled with 'acceptable' or 'good' statistical power.

Step 3. Sort background into ascending order and set the upper prediction limit equal to a large background order statistic (*e.g.*, the maximum or second largest value). If all constituent measurements in a background sample are non-detect, use the Double Quantification rule in **Chapter 6**. The constituent should not be included in calculations identifying the target false positive rate.

Step 4. Collect two initial measurements per compliance well. If both do not exceed the upper prediction limit, the test passes since the median of order 3 will also not exceed the limit. There is no need to collect the third initial observation or any resamples. If both exceed the prediction limit, the median will also exceed the limit. There is no need to collect the third initial measurement. If using a 1-of-1 plan, move to Step 5. Otherwise, collect up to three resamples in order to assess the resample median.

If one initial measurement is above and one below the limit, collect a third observation to determine the position of the median relative to the prediction limit. In all cases, if two or more of the compliance point observations are non-detect, set the median equal to the quantification level (QL).

Step 5. Compare the median value for each compliance well against the upper prediction limit. If a 1-of-2 retesting scheme is selected and any compliance point median exceeds the limit, collect up to three additional resamples from that well. Compute the resample median and compare this value to the upper prediction limit.

Identify a compliance well as potentially contaminated when either the initial median exceeds the upper prediction limit for a 1-of-1 plan, or both the initial median and the resample median exceed the prediction limit in a 1-of-2 plan. Conversely, declare a well to have passed the test if the initial median does not exceed the prediction limit, or the resample median in a 1-of-2 scheme does not exceed it.

## ►EXAMPLE 19-5

The following trace mercury data have been collected in the past year from a site with four background wells and 10 compliance wells (two of which are shown below). The facility must monitor for five constituents, including mercury. Assuming that the percentage of non-detects in background is too high to make a parametric analysis appropriate or feasible, compare interwell non-parametric prediction limits for both observations and medians at the annual statistical evaluation, and determine whether either compliance well indicates significant evidence of mercury contamination. Further assume that the sequentially reported compliance well data below are obtained as needed for the different test comparisons.

| | **Mercury Concentrations (ppb)** | | | | | |
|---|---|---|---|---|---|---|
| Event | BG-1 | BG-2 | BG-3 | BG-4 | CW-1 | CW-2 |
| 1 | .21 | <.2 | <.2 | <.2 | .22 | .36 |
| 2 | <.2 | <.2 | .23 | .25 | .20 | .41 |
| 3 | <.2 | <.2 | <.2 | .28 | <.2 | .28 |
| 4 | <.2 | .21 | .23 | <.2 | .25 | .45 |
| 5 | <.2 | <.2 | .24 | <.2 | .24 | .43 |
| 6 | | | | | <.2 | .54 |

### SOLUTION

Step 1. Using a target SWFPR of 10%, compute the target per-constituent false positive rate, noting that the monitoring list consists of five parameters. This implies that $\alpha_{const} = 1 - (1-.1)^{1/5} = .021$ using equation [19.16]. Since the detection rate in background is only 35%, it is reasonable to consider non-parametric prediction limits with retesting. The background sample size of $n = 20$ is to be used to construct an interwell prediction limit for all $w = 10$ compliance wells. Since there is only one annual evaluation ($n_E = 1$), the look-up table margin value of $w^*$ equals $w \times n_E = 10$.

Step 2. Determine potentially applicable retesting plans. First consider non-parametric prediction limits for individual observations with $n = 20$ and $w = 10$. Consulting **Tables 19-19** through **19-22** in **Appendix D**, only the 1-of-3, 1-of-4, and modified California plans meet (*i.e.*, do not exceed) the target false positive rate of 2.1%. To use the 1-of-3 or modified California plans, the prediction limit needs to be set to the maximum background measurement. In the 1-of-4 plan, the prediction limit can be set to either the maximum or second-highest value in background using the **Appendix D** tables. A final 1-of-4 plan determined with the *Optimal Rank Values Calculator* allows the use of the 3rd highest value. All of these plans boast good power compared to the annual ERPC. Both the 1-of-4 and modified California schemes may require as many as 3 separate and independent resamples in addition to the initial observation.

Consider tests for future medians of order 3 in **Tables 19-23** and **19-24** in **Appendix D**. Only the 1-of-2 plan using the maximum background value as the prediction limit meets the $\alpha_{const}$ target. It also has good power, but requires 3 initial measurements and up to 3 additional individual resamples.

Step 3. Sort the combined background data and compute the possible prediction limits as $PL_{(n)}$ = .28 ppb, $PL_{(n-1)}$ = .25 ppb, and $PL_{(n-2)}$ = .24 ppb, respectively representing the maximum, second-largest, and third-largest background values.

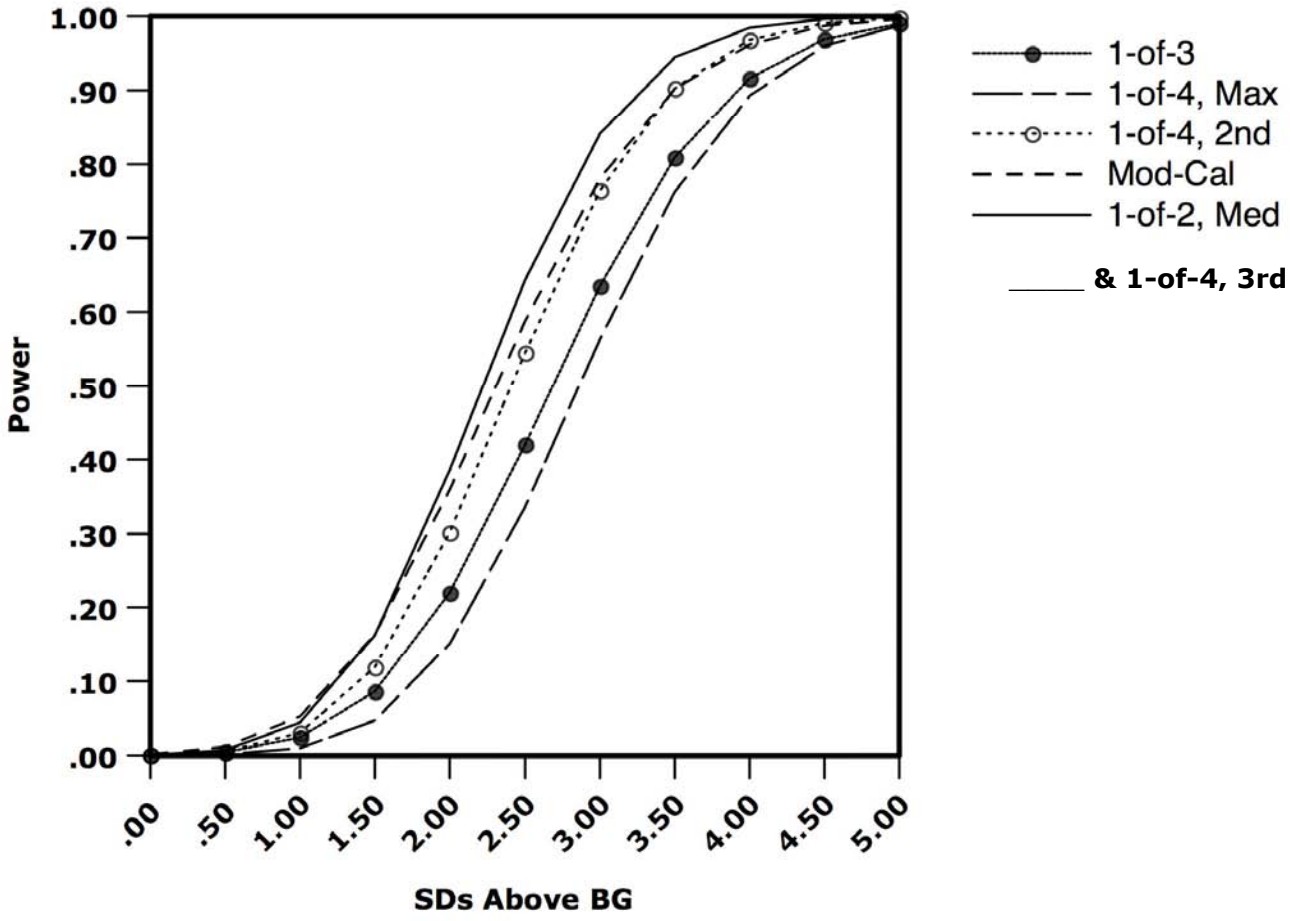
Step 4. Determine the test outcomes at each compliance well using the various retesting plans, as shown in the table below. For the prediction limits on individual observations, the first sample collected during Event 1 is used as the initial screen to determine if any resampling is necessary. The first 3 measurements at each compliance well are used to form the initial comparison. The median at CW-1 is .20 ppb, while that at CW-2 is .41 ppb.

| Compliance Well | Retesting Plan | Achieved α | # Initial Samples | Resamples Required | BG Limit | Result |
|---|---|---|---|---|---|---|
| CW-1 | 1-of-3 | .0055 | 1 | 0 | .28 | Pass |
| | 1-of-4, Max | .0009 | 1 | 0 | .28 | Pass |
| | 1-of-4, 2nd | .0046 | 1 | 0 | .25 | Pass |
| | 1-of-4, 3rd | .0135 | 1 | 0 | .24 | Pass |
| | Mod-Cal | .0140 | 1 | 0 | .28 | Pass |
| | 1-of-2, Med | .0060 | 3 | 0 | .28 | Pass |
| CW-2 | 1-of-3 | .0055 | 1 | 2 | .28 | Pass |
| | 1-of-4, Max | .0009 | 1 | 2 | .28 | Pass |
| | 1-of-4, 2nd | .0046 | 1 | 3 | .25 | Fail |
| | 1-of-4, 3rd | .0135 | 1 | 3 | .24 | Fail |
| | Mod-Cal | .0140 | 1 | 3 | .28 | Fail |
| | 1-of-2, Med | .0060 | 3 | 3 | .28 | Fail |

All of the acceptable plans indicate that CW-1 is not statistically different from background, although more initial sampling is required for the 1-of-2 retesting plan with medians. For CW-2, the results are more problematic. The 1-of-3 and 1-of-4 plans based on the maximum background value allow the well to pass, while the other four plans indicate a significant difference from background. The least degree of sampling is required by the 1-of-3 plan; at some facilities, greater sampling efforts may not be feasible. When a well is likely to be contaminated, the number of samples required to actually make a decision about the well is similar across the plans with the exception of the 1-of-2 prediction limit on a median.

A further consideration is that although the power of each plan exceeds the annual ERPC when additional resampling is possible, it is helpful to compare the full power curves of multiple plans to determine whether a particular plan offers greater power than the rest. **Figure 19-2** displays an overlay of the six power curves associated with the retesting plans in this example. For these inputs, the 1-of-2 retesting plan for a median of order 3 using the background maximum and the 1-of-4 plan on individual observations using the 3rd highest background value achieve the best overall power (shown as a single curve on **Figure 19-2**).

Figure 19-2. Comparison of Full Power Curves




As seen in **Figure 19-2**, the two plans that pass the second compliance well have visibly lower power — especially in the range of 2 to 3.5 standard deviations above background — than the four plans that failed CW-2. In such a situation, the user needs to carefully balance the risks and benefits of each acceptable resampling plan. In some cases, the cost of greater amounts of resampling may be outweighed by the added sensitivity of the test to evidence of groundwater contamination. ◄

*This page intentionally left blank*

# CHAPTER 20. MULTIPLE COMPARISONS USING CONTROL CHARTS



This chapter describes control charts, a second recommended core strategy for detection monitoring. Control charts are a useful and powerful alternative to prediction limits. The Unified Guidance is the first EPA document to discuss retesting and simultaneous testing of multiple wells and/or constituents as they relate to control charts. Research of these topics is still ongoing.

## 20.1 INTRODUCTION TO CONTROL CHARTS

Control charts are a viable alternative to parametric prediction limits for testing groundwater in detection monitoring. They are similar to prediction limits for future observations in that a control chart limit is estimated from background and then compared to a sequence of compliance point measurements. If any of these values exceeds the control limit, there is initial evidence that the compliance point concentrations exceed background.

Control charts can be constructed as either interwell or intrawell tests. The main difference is how background is defined and what measurements are utilized to build the control limit. Interwell control charts establish the control limit from designated upgradient and potentially other background wells. Intrawell control charts, on the other hand, employ historical measurements from a compliance point well as background. Intrawell tests can only be appropriately applied if the historical compliance well background is uncontaminated.

An advantage of control charts over prediction limits is that a control chart graphs the compliance data over time. Certain varieties can also evaluate gradual increases above background over the period of monitoring. Trends and changes in concentration levels can be easily seen since the sample observations are consecutively plotted on the chart. This provides the analyst an historical overview of the pattern of measurement levels. Prediction limits are typically constructed to allow only ***point-in-time comparisons*** between the most recent compliance data and background, making long-term trends more difficult to identify.[1]

[1] Long-term results from repeated application of a prediction limit can be plotted over time, creating a graph similar in nature to a control chart. But this has been infrequently done in practice.

As a well-established statistical methodology, there are many kinds of control charts. Historically, control charts have been put to great use in quality engineering and manufacturing, but have more recently been adapted for use in groundwater monitoring. The specific control chart recommended in the Unified Guidance is known as a combined Shewhart-CUSUM control chart (Lucas, 1982). It is a 'combined' chart because it simultaneously utilizes two separate control chart evaluation procedures. The Shewhart portion is almost identical to a prediction limit in that compliance measurements are individually compared against a background limit. The cumulative sum [CUSUM] portion sequentially analyzes each new measurement with prior compliance data. Both portions are used to assess the similarity of compliance data to background in detection monitoring.

The Shewhart-CUSUM control chart works as follows. Appropriate background data are first collected from the specific compliance well for intrawell comparisons or from separate background wells for interwell tests. The baseline parameters for the chart, estimates of the mean and standard deviation, are obtained from these background data. These baseline measurements characterize the expected background concentrations at compliance wells.

As future compliance observations are collected, the baseline parameters are used to standardize the newly gathered data. After these measurements are standardized and plotted, a control chart is declared *out-of-control* if future concentrations exceed the baseline control limit. This is indicated on the control chart when either the Shewhart or CUSUM plot traces begins to exceed a control limit. The limit is based on the rationale that if the well remains uncontaminated as it was during the baseline period, new standardized observations should not deviate substantially from the baseline mean. If a release occurs, the standardized values will deviate significantly from baseline and tend to exceed the control limit. The historical baseline parameters then no longer accurately represent current well concentration levels.

Combined Shewhart-CUSUM control charts initially featured two control limits, one for testing the Shewhart portion of the chart, one for testing the CUSUM portion of the chart. Later research on control charts (Davis, 1999; Gibbons, 1999) indicated that having separate control limits for the Shewhart and CUSUM procedures is generally not important. Both control chart traces can instead be compared to a ***single*** control limit. This modification not only makes the control chart method slightly easier to apply, but also aids in measuring the statistical performance of control charts over a variety of monitoring networks.

## 20.2 BASIC PROCEDURE

The basic procedure for constructing a control chart is presented below. Requirements and assumptions for control charts are discussed in later sections:

Step 1. Given $n$ background measurements ($x_{jB}$), estimate the baseline parameters by computing the sample mean ($\bar{x}_B$) and standard deviation ($s_B$).

Step 2. For a compliance point measurement ($x_i$) collected on sampling event $T_i$, compute the standardized concentration $Z_i$:

$$Z_i = \left(x_i - \bar{x}_B\right) \Big/ s_B \qquad [20.1]$$

Step 3. For each sampling event $T_i$, use the standardized concentrations from **Step 2** to compute the standardized CUSUM $S_i$. Set $S_0 = 0$ when computing the first CUSUM $S_1$.

$$S_i = \max\left[0,\ \left(Z_i - k\right) + S_{i-1}\right] \qquad [20.2]$$

The notation max[*A*, *B*] in equation [20.2] refers to picking the maximum of quantities *A* and *B*. Furthermore, the parameter *k* designates half the ***displacement*** or shift in standard deviations that should be quickly detected on a control chart. Often *k* is set equal to 1, meaning that the control chart will be designed to rapidly detect upward concentration shifts of at least two standard deviations. Since $Z_i$ is standardized by the estimated baseline standard deviation, an increase of *r* units in $Z_i$ corresponds to an increase of *r* standard deviations above the baseline mean in the domain of concentrations $x_i$.

Step 4. To plot the control chart in concentration units, compute the ***non-standardized*** CUSUMs $S_i^c$ with the equation:

$$S_i^c = \bar{x}_B + S_i \cdot s_B \qquad [20.3]$$

Step 5. Calculate the non-standardized control limit used to assess compliance of both future measurements ($x_i$) and non-standardized CUSUMs ($U_i$). Traditionally, two parameters were used to compute standardized limits: the decision internal value ($h$) and the Shewhart Control Limit ($SCL$). The Unified Guidance instead recommends only one standardized control limit ($h$). Compute the non-standardized control limit ($h_c$) as:

$$h_c = \bar{x}_B + h \cdot s_B \qquad [20.4]$$

Step 6. Construct the control chart by plotting both the compliance measurements ($x_i$) and the non-standardized CUSUMs ($S_i^c$) on the *y*-axis against the sampling events $T_i$ along the *x*-axis. Also draw a horizontal line at the concentration value equal to the control limit, $h_c$.

Step 7. Moving forward in time from the first plotted sampling event $T_1$, declare the control chart to be potentially out-of-control if either of two situations occurs: 1) the trace of non-standardized concentrations exceeds $h_c$; or 2) the CUSUMs become too large, exceeding $h_c$.

The first case signifies a rapid increase in concentration level among the most recent sample data. The second can represent either a sudden rise in concentration levels or a gradual increase over time. A gradual increase or trend is particularly indicated if the CUSUM exceeds the control limit but the compliance concentrations do not. The reason for this is that several consecutive, small, increases in $x_i$ will not trigger the control limit, but may cause a large enough increase in the CUSUM. As such, a control chart can indicate the onset of either sudden or gradual contamination at the compliance point.

## ►EXAMPLE 20-1

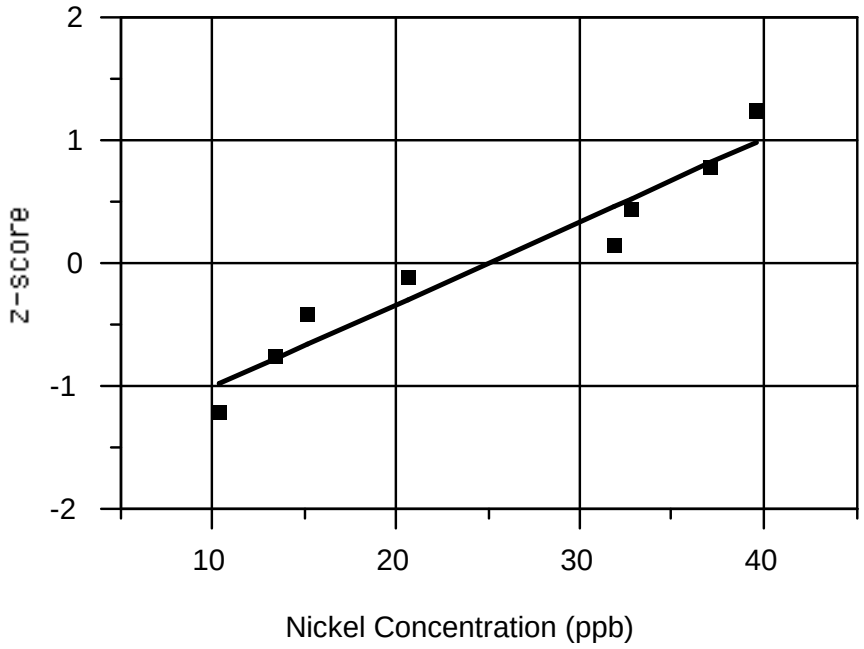
For background nickel data collected during 8 months in 1995 shown below, construct an intrawell control chart and compare it with the first 8 months of the compliance period (1996):

| Month | Nickel Concentration (ppb) | |
|---|---|---|
| | Baseline Period (1995) | Compliance Period (1996) |
| 1 | 32.8 | 19.0 |
| 2 | 15.2 | 34.5 |
| 3 | 13.5 | 17.8 |
| 4 | 39.6 | 23.6 |
| 5 | 37.1 | 34.8 |
| 6 | 10.4 | 28.8 |
| 7 | 31.9 | 23.7 |
| 8 | 20.6 | 81.8 |

### SOLUTION

Step 1. As discussed in **Section 20.3.3**, control charts are a parametric procedure requiring normal or normalized data. Test the $n$ = 8 baseline measurements for normality. A probability plot of these data provided in **Figure 20-1** exhibits a mostly linear trend. The Shapiro-Wilk test statistic computed for these data is $W = 0.896$. Compared to the $\alpha = .10$ level critical point of $w_{.10,8} = 0.851$ (**Table 10-3** of **Appendix D**), the Shapiro-Wilk test indicates that the baseline data are approximately normal. Construct the control chart using the original nickel measurements.

Figure 20-1. Probability Plot of Baseline Nickel Data



Step 2. Use the 1995 baseline nickel data to compute the sample mean and standard deviation: $\bar{x}_B$ = 25.14 ppb and $s_B$ = 11.518 ppb. Then compute the standardized concentration $Z_i$ for each 1996 compliance period sampling event using equation [20.1]. These values are listed in the fourth column of the table below.

| Month | $T_i$ | Nickel (ppb) | $Z_i$ | $Z_i - k$ | $S_i$ | $S_i^c$ |
|---|---|---|---|---|---|---|
| 1 | 1 | 19.0 | −0.53 | −1.53 | 0.00 | 25.14 |
| 2 | 2 | 34.5 | 0.81 | −0.19 | 0.00 | 25.14 |
| 3 | 3 | 17.8 | −0.64 | −1.64 | 0.00 | 25.14 |
| 4 | 4 | 23.6 | −0.13 | −1.13 | 0.00 | 25.14 |
| 5 | 5 | 34.8 | 0.84 | −0.16 | 0.00 | 25.14 |
| 6 | 6 | 28.8 | 0.32 | −0.68 | 0.00 | 25.14 |
| 7 | 7 | 43.7 | 1.61 | 0.61 | 0.61 | 32.16 |
| 8 | 8 | 81.8 | 4.92 | 3.92 | 4.53 | 77.31 |

Step 3. Compute the standardized CUSUMs as follows. First let the shift displacement parameter $k$ = 1 and set $S_0 = 0$. After subtracting k from each $Z_i$, calculate the CUSUM using equation [20.2] . Note that none of the CUSUMs are positive until the first occurrence of a positive quantity ($Z_i - k$). As shown in the sixth column above, the standardized CUSUMs for the 6th, 7th and 8th events are calculated as:

$$S_6 = \max\left[0, \left(0.32 - 1\right) + 0\right] = 0$$
$$S_7 = \max\left[0, \left(1.61 - 1\right) + 0\right] = 0.61$$
$$S_8 = \max\left[0, \left(4.92 - 1\right) + 0.61\right] = 4.53$$

Step 4. Calculate the non-standardized CUSUMs ($S_i^c$) using the individual $Z_i$, baseline mean and standard deviation parameters in equation [20.3]. These values are listed in the last column of the table above. For the 8th sampling event, this calculation gives:

$$S_8^c = 25.14 + 11.518\left(4.53\right) = 77.31$$

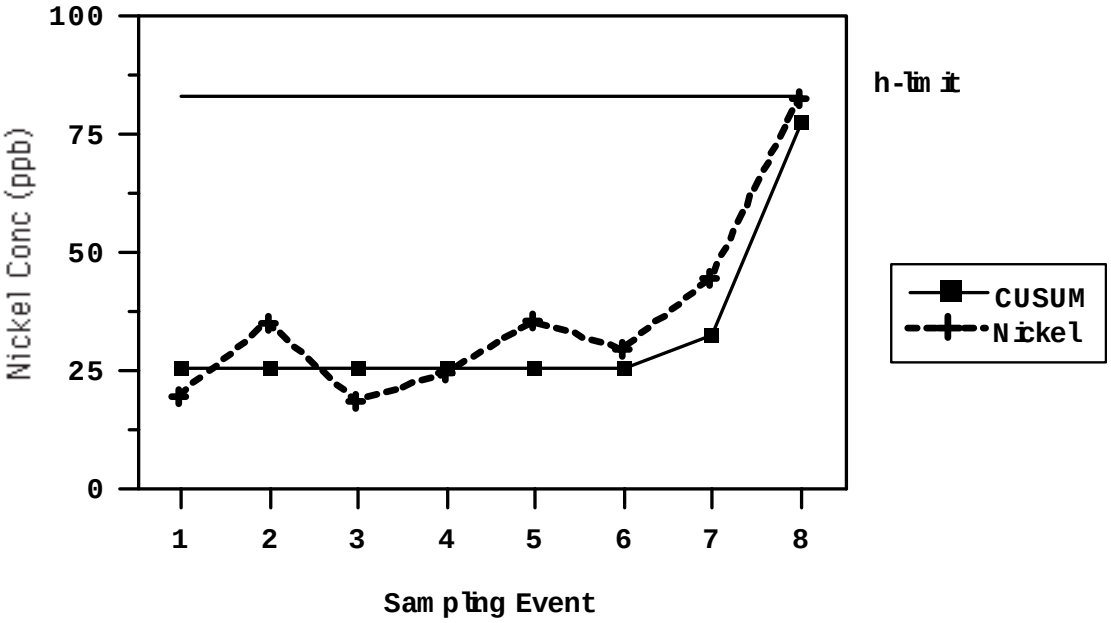
Step 5. Compute the non-standardized control limit using equation [20.4]. For purposes of this example, set $h$ = 5; the non-standardized limit becomes:

$$h_c = 25.14 + 11.518\left(5\right) = 82.73 \text{ ppb}$$

Step 6. Using the compliance period nickel concentrations and the non-standardized CUSUMs, plot the control chart as in **Figure 20-2**. The combined chart indicates there is insufficient evidence of groundwater contamination in 1996 because neither the nickel concentrations nor the CUSUM statistics exceed the control limit for the months examined. However, both traces nearly exceed $h_c$, and conceivably might do so in future sampling events if the apparent trend continues. If that were to happen, retesting can be performed to better determine whether the

increase was one or a series of chance fluctuations or an actual mean-level change in nickel concentrations. ◄

Figure 20-2. Shewhart-CUSUM Control Chart for Nickel Measurements




## 20.3 CONTROL CHART REQUIREMENTS AND ASSUMPTIONS

As with other statistical methods, control charts are based on certain assumptions about the sample data. There are also some minimum requirements for constructing them. None of the assumptions or requirements are unique to control charts, although there are some special issues.

### 20.3.1 STATISTICAL INDEPENDENCE AND STATIONARITY

The methodology for control charts assumes that the sample data are statistically independent. A control chart can give misleading results if consecutive sample measurements are serially correlated (*i.e.*, autocorrelated). For this reason, it is important to design a sampling plan so that distinct volumes of groundwater are analyzed at each sampling event (**Section 14.3.1**). Duplicate laboratory analyses (*i.e.*, aliquot or field splits) should also not be treated as independent observations when constructing a control chart. Gibbons (1999) recommends that control chart observations be collected no more frequently than quarterly. Since physical independence does generally not guarantee statistical independence (**Section 14.1**), a test of autocorrelation using the sample autocorrelation function or rank von Neumann ratio tests (**Section 14.2**) should be performed to determine whether the current sampling interval affords uncorrelated measurements.

If the background data exhibit a clear seasonal cyclical pattern, the values should be deseasonalized before computing the control chart baseline parameters. For a seasonal pattern at a single well, the

method of **Section 14.3.3.1** can be used to create adjusted measurements having a stable mean. At several or a group of wells indicating a common seasonal pattern, the adjusted values can be computed using a one-way analysis of variance [ANOVA] for temporal effects (**Section 14.3.3.2**). When baseline data are deseasonalized, it is essential that newly collected compliance measurements also be deseasonalized in the same manner. It is presumed that the same pattern or physical cause will impact future data in the same manner as for the baseline measurements.

To deseasonalize compliance point measurements, simply use the seasonal and grand means estimated from background in computing the adjusted compliance point values. If the control chart remains in control following deseasonalizing, the existing background can be updated with the newer measurements. However, the revised background set should be checked again for seasonality and the seasonal and grand means re-computed, in order to more accurately adjust future measurements.

Control charts also assume that the background mean is *stationary over time*. This means there should be no apparent upward or downward trend in the background measurements. A trend imparts greater-than-expected variation to the background data, increasing the baseline standard deviation and ultimately the control limit. The net result is a control chart that has less power to identify groundwater contamination. Tests for trend described in **Chapter 17** can be used to check the assumption of no background trends. Should an upward or downward trend be verified, *the background data should not be de-trended*. While it is possible to construct and use a control chart with de-trended background and future data, the assumption that the trend will continue indefinitely is very problematic. The trend should first be investigated to ensure that background has been properly designated. Other monitoring wells should be checked to see if the same trend is occurring, indicating either evidence of an earlier release or possibly a sitewide change in the aquifer. In any case, a switch should be made to a trend test rather than a control chart.

As noted, control charts can be employed as either interwell or intrawell tests. However, interwell control charts require a spatially stationary mean across the monitoring network. If spatial variability exists among background wells for certain constituents, interwell control charts will be no more interpretable than prediction limits. A related problem can plague intrawell control charts if there is *prior* spatial variability (*i.e.*, some compliance wells are already contaminated prior to selection of intrawell backgrounds). *Historical observations should be used as baseline data in intrawell tests only if the compliance wells are known to be unaffected by a release from the monitored unit.* Otherwise, the control limit based on the greater-than-expected background values may be set too high to identify current contamination.

## 20.3.2 SAMPLE SIZES AND UPDATING BACKGROUND

Both background mean and standard deviation estimates are needed to construct a control chart limit. The Unified Guidance recommends at least $n = 8$ measurements for the defining the baseline, particularly to ensure an accurate standard deviation estimate. *Baseline* observations are traditionally not plotted on the chart, although it may be visually helpful to include background values on the plot using a distinct symbol (*e.g.*, hollow instead of filled symbol).

Whether baseline observations are obtained from upgradient background wells for interwell testing or from individual compliance well historical data for intrawell use, these data are only small random samples used to estimate the true background population characteristics. Any particular sample set may not be adequately representative. Because of this likelihood, the background sample size requirements suggested above for constructing a control chart should be regarded as a minimum. More background observations should preferably be added to the initial set to improve the characterization of the background distribution.

For interwell control charts, periodic updating of background (**Chapter 5**) poses no difficulty. New observations should be collected at background wells on each sampling event. Then, every 1-2 years, the newly collected background should be added to the existing background pool after testing/checking for statistical similarity. The revised background can be used to re-compute the baseline parameters and, in turn, the control limit.

Updating background for intrawell control charts depends on the control chart remaining 'in-control' for several consecutive sampling events. As long as a confirmed exceedance does not occur, the in-control compliance measurements collected since the last background update can be tested against the existing background for statistical similarity using a Student's *t*- or Wilcoxon rank-sum test (**Section 5.3**). ASTM Standard D6312-98 (1999) recommends testing the newly revised background set for trends, using trend tests including those in **Chapter 17**. The ASTM methodology is intended to avoid incorporating a subtle trend into the control chart background, which influences the re-computed baseline parameters and weakens the statistical power of the control chart to identify contaminant releases.

If the comparison of recent in-control measurements against existing background indicates a statistically significant difference, it may reflect changes in natural groundwater conditions unrelated to contamination events. In these circumstances, it is possible to update background by creating a 'moving window.' The background sample size $n$ remains fixed, with only the most recent $n$ measurements included as background for computing baseline parameters. Earlier sampling events are excluded. The overriding goal is to ensure that background reflects the most current and representative groundwater conditions (**Chapter 3**).

Despite the apparent benefits, the statistical performance of control charts is only partially known when background is periodically updated. Davis (1999) has performed the most extensive simulations of this question. He suggests that substantially different simulation results occur with the CUSUM portion when background is periodically updated (especially early on) and combined with either a small maximum *run length* or a 'warm-up' period or both (see **Section 20.4.1**).

Two other issues affect both control charts and prediction limits when updating intrawell background. First, if background is periodically augmented by adding new measurements (either from upgradient background wells or from recent in-control compliance measurements), the overall background sample size is increased. This in turn should cause the prediction or control chart limit to decrease.

For instance, prediction limit tables in **Chapter 19** demonstrate that as the background sample size increases, lower prediction limit $\kappa$–multipliers are appropriate. The expanded background sample is used to re-compute the prediction limit, provided that the measurements added to background do not indicate an adverse change in groundwater quality. New compliance measurements are then tested against the revised prediction limit. But the same cannot be done with control charts unless the CUSUM is *reset to zero*. The reason is that the CUSUM will have *already been affected* by those compliance measurements now being added to intrawell background. An independent comparison between compliance point values and background is thus precluded. Consequently, the Unified Guidance recommends that the CUSUM portion of the control chart be reset after each periodic update of intrawell background.[2]

The second issue is how to update intrawell background when an initial measurement has exceeded the control or prediction limit, but one or more resamples disconfirm the exceedance. Routine detection monitoring continues in this situation. No confirmed exceedance is registered for a prediction limit test and the control chart remains in-control. Should the initial exceedance be included or excluded when later updating intrawell background?

The Unified Guidance recommends a strategy parallel to the handling of outliers (**Chapter 12**). If the exceedance can be shown to be a measurement in error or a confirmed outlier, it should be excluded from the revised background. Otherwise, any disconfirmed exceedances (including any resamples that exceed the background limit but are disconfirmed by other resamples) should probably be included when updating the background. The reason is that background limits designed to incorporate retesting are computed as low as possible to ensure adequate statistical power. The trade-off is that compliance measurements legitimately similar to background but drawn from the upper tail of the distribution, sometimes exceed the limit and have to be disconfirmed with a resample. Any exceedance not documented as an error or outlier is most likely representative of some portion of the background population that previously had gone unsampled or unobserved.

### 20.3.3 NORMALITY AND NON-DETECT DATA

The combined Shewhart-CUSUM control chart is a parametric procedure. This implies that background used to estimate the baseline parameters should either be normal or normalized via a transformation. Normality can be tested on either the raw measurement or transformed scale using one of the goodness-of-fit techniques described in **Chapter 10**. If the hypothesis of normality is accepted,

[2] The same 'overlapping' dependence between the CUSUM and revised background will also be true when background is updated using a 'moving window' approach. The CUSUM should therefore be reset in these cases too. However, since the background sample size is kept fixed, the standardized control limit ($h$) will not decrease as it does when background is augmented.

construct the control chart on the raw measurements. If it is rejected, try a transformation and retest the transformed data for normality. If the transformation works to normalize background, construct the control chart on the transformed measurements, being sure to use the same transformation on both background and the compliance values to be plotted.

Unlike prediction limits, no non-parametric version of the combined Shewhart-CUSUM control chart exists. If the background sample cannot be normalized perhaps due to a large fraction of non-detects, a non-parametric prediction limit should be considered (**Section 19.4**). Control charts will be most appropriate for those constituents with a reasonably high detection frequency. These include many inorganic constituents (***e.g.***, certain trace elements, indicators and geochemical monitoring parameters) that occur naturally in groundwater, or for other persistently detected, site-specific organic chemicals.

If no more than 10-15% of the data are non-detect, it may be possible to normalize the data via simple substitution (**Section 15.2**) of half the reporting limit [RL] for each background non-detect. A normalizing transformation can sometimes be found using a *censored probability plot* (**Chapter 15**) for background data containing a substantial fraction of non-detects up to 50%. A censored estimation technique such as Kaplan-Meier or Robust Regression on Order Statistics [Robust ROS] (**Chapter 15**) can then be used to compute estimates of the baseline mean ($\hat{\mu}_B$) and standard deviation ($\hat{\sigma}_B$) that account for the left-censored measurements. These adjusted estimates should replace the background sample mean ($\overline{x}_B$) and standard deviation ($s_B$) in the control chart equations of **Section 20.2**. The Unified Guidance differs somewhat from the recommended approach in ASTM Standard D6312-98 (ASTM, 1999), which is to set all non-detects identically to zero.

No matter how background non-detects are treated, control charts require an additional step for future observations that isn't needed with prediction limits. Each new compliance point measurement statistic must be added to the CUSUM associated with previous sampling events. If the new observation is a non-detect, some value (typically a fraction of the RL) needs to be imputed for the censored measurement in order to update the CUSUM. The Unified Guidance recommends that half the RL be substituted for these measurements.[3]

[3] If an intrawell control chart is constructed and it remains 'in-control' until the next background update, any non-detects observed in the meantime should be treated as left-censored measurements for purposes of updating the baseline mean and standard deviation estimates. In other words, the simple substitution of RL/2 should only apply temporarily to compute an updated CUSUM.

## 20.4 CONTROL CHART PERFORMANCE CRITERIA

A significant difference exists between control charts and prediction limits in setting statistical performance criteria. Standard equations described in previous chapters allow the user to generate an exact confidence level (1–α) for prediction limits. Obtaining similar confidence levels for the Shewhart-CUSUM control charts needs to be done experimentally through varying the two background control chart limits ($h$) and the displacement parameter ($k$), as well as the retesting options. The control chart parameter limits in the two previous EPA RCRA statistical guidance documents were based on work by Lucas (1982), Hockman & Lucas (1987), and Starks (1988). Monte Carlo simulations for various combinations of control chart parameters (without retests) were used to develop the overall recommendations in their papers.

The specific parameter choices were not fixed, but appeared to work best in simulations at a single well. Starks (1988) recommended setting $h = 5$ and $k = 1$ for standardized measurements, especially in the early stages of monitoring. He further suggested that after 12 consecutive in-control measurements, the baseline mean and standard deviation be updated to include more recent sampling measurements. The values of $k$ and *SCL* (the separate Shewhart control limit) could then be reduced to $k = 0.75$ and *SCL* = 4.0. In effect, this tightens the control chart limits to reflect that additional data are available to better characterize the baseline population.

More recent research (notably Gibbons, 1999) has demonstrated that control charts from the quality control literature do not account for several important characteristics of groundwater monitoring networks. The most important is the problem of multiple comparisons (*i.e.*, the need to simultaneously conduct testing of many well-constituent pairs during an evaluation period described in **Chapter 6**). Control chart performance is typically assessed on an individual well basis, rather than over a network of simultaneous tests. The recommended control limits have no obvious connection to the expected false positive rate (α), nor is the traditional control limit adjustable like the κ-factor in prediction limits. There is a need to account for differences in background sample sizes, a desired false positive rate, and the number of monitoring network tests in similar fashion to prediction limits. Moreover, early research and guidance did not address the issue of retesting in control charts. Retesting provides substantial improvements in prediction limit performance, and its potential needs to be evaluated for control charts.

It is standard practice to discuss the performance of prediction limits in terms of statistical power and false positive rates. However, statistical performance of control charts is usually measured via the *average run length* [ARL]. The ARL is the average number of sampling events before the control limit is first exceeded, identifying an 'out-of-control' process. Ideally, the ARL should be large when the mean concentration of the tested constituent is at or near the baseline average, but increasingly smaller as the true mean is gradually shifted above baseline.

Put in standard statistical terms, the control chart should not easily or quickly signal false evidence of a release when a release has *not* occurred. To have a low false positive rate when the null hypothesis of no contamination is true, the chart should stay 'in-control' for a long time indicated by a large ARL. The statistical power for detecting a release when it occurs should be as high as possible. A short ARL will indicate that a control chart is quickly determined to be out-of-control.

False positive rates ($\alpha$) for CUSUM control charts cannot be equated precisely with ARLs. But it has been found that the ARLs closely follow a geometric distribution pattern with a mean equal to ($1/\alpha$). Thus, a control chart with an ARL of 100 would have an associated false positive rate of roughly 1%. The relationship is not exact, especially for combined Shewhart-CUSUM control charts. It is also affected by the randomness in the background data used to establish the control chart baseline.

Thus, the Unified Guidance offers a new framework for measuring control chart statistical performance. It is suggested that measuring false positive rates in control charts be conducted by establishing a time frame or run length of interest, specifically, a period of one year. A false positive is counted if the chart has a confirmed exceedance sometime during the year, under the assumption of no contaminant release. Statistical power is similarly evaluated for a fixed time interval (***e.g.***, one year) by measuring the proportion of run lengths with confirmed exceedances ***during that interval***. In this way, both the false positive rate and power are tied to a specific one-year time frame.

This framework is consistent with the guidance recommendations that prediction limit performance be measured according to an annual, cumulative 10% site-wide false positive rate [SWFPR] and that cumulative, annual effective power be comparable to the EPA reference power curves [ERPC]. The suggested framework for control charts allows a direct comparison with prediction limits when designing alternate statistical approaches.

### 20.4.1 CONTROL CHARTS WITH MULTIPLE COMPARISONS

Until recently, control charts were not designed to address the SWFPR when testing multiple well-constituent pairs. Furthermore, it was not clear to a user how to adjust for multiple tests using fixed control limits (SCL, $k$ and $h$). Because of these problems, Gibbons (1999) performed a series of Monte Carlo simulations to gauge intrawell control chart performance for up to 500 simultaneous tests. Gibbons also examined the outcomes when the single Shewhart and CUSUM decision limit was allowed to vary between $h$ = {4.5, 5.0, 5.5, and 6.0}. He found that control charts could be designed with both high power and a low SWFPR, as long as retesting was incorporated into the methodology.

Additional Monte Carlo simulation work was performed by Davis (1999). He found that control charts perform similarly to prediction limits when both use retests. But he also noted that certain favorable outcomes in Gibbons (1999) were the result of combining frequent updating of background and a 'warm-up' period for the chart. In the latter period, any control limit exceedances were ignored. The simulations were based on small ***maximum run lengths.***

Other researchers have noted (for instance, Luceño and Puig-Pey, 2000) that the run length distribution of CUSUM control charts is often close to geometric. This implies that even when the ARL is large, there can be significant probability of an early failure. The difficulty in a real-life setting is that one will not know whether an early exceedance of the control limit is due to contaminated groundwater or simply a false positive exceedance for an otherwise in-control chart. This guidance recommends against the use of 'warm-up' periods when implementing or assessing the performance of Shewhart-CUSUM control charts.

Gibbons (1999) provides results for a number of control chart limit options, but does not determine limits which can provide exact false positive rate control. A number of potential commonly applied retesting strategies are also not evaluated. In contrast, both Gibbons (1994) and the Unified Guidance (**Chapter 19**) do provide such control for prediction limits using a wider array of retesting strategies.

Facilities may need to conduct theirown specific Monte Carlo simulations if the published literature options cannot be applied at their site. Simulations might be needed for either intrawell or interwell control charts or both. Overall methodologies for Monte Carlo simulations are provided below. The first step for either type test is a simulation of the cumulative annual false positive rate. Then a second simulation measures the cumulative, annual statistical power.

To perform an *intrawell* simulation, repeat the following steps for a large number of simulations (*e.g.*, $N_{sim}$ = 10,000):

1. Determine the total number of well-constituent pairs for which statistical testing is required, as well as the number of pairs at which intrawell control charts will be constructed. Use the basic subdivision principle (**Section 19.2.1**) to determine the per-test false positive rate ($\alpha_{test}$) associated with each control chart that meets the target SWFPR.

2. Determine the intrawell background sample size ($n$). Generate $n$ standard normal measurements. Then form baseline estimates by computing the sample mean ($\bar{x}_B$) and standard deviation ($s_B$).

3. Pick a set of possible standardized control limits ($h$). Choose a maximum run length ($M$), based on the number of sampling events conducted each year (*e.g.*, $M$ = 4 for quarterly sampling).

4. For each potential control limit ($h$), compute the non-standardized control limit using equation [20.4]. Then simulate the behavior of the control chart from sampling event 1 to sampling event $M$ by generating standard normal compliance measurements for each event. Generate enough random measurements to account for resamples potentially needed with a selected retesting strategy.

5. Test the initial measurement associated with each sampling event against the non-standardized control limit. Also form the CUSUM for events 1 to $M$ using equations [20.2] and [20.3]. Compare the non-standardized CUSUM against the control limit.

6. If either the initial measurement or the CUSUM exceeds $h_c$, use the resample(s) for that sampling event to perform a retest (see below). If the retest confirms the initial exceedance, record a false positive for that particular simulation (out of $N_{sim}$).

7. After all $N_{sim}$ runs have been conducted, compute the observed false positive rate ($\alpha_h$) associated with each possible *standardized* control limit ($h$) by dividing $N_{sim}$ into the number of observed false positives. Set the final control limit equal to that value of $h$ for which $\alpha_h$ is closest to $\alpha_{test}$.

The simulation for an *interwell* control chart is similar to the intrawell case, with a few key differences. First, instead of a per-test false positive rate, the basic subdivision principle must be used to compute a *per-constituent* false positive rate ($\alpha_{const}$). The reason is that the same background measurements for a given constituent are used to test each of the compliance wells in the network.

Secondly, when generating standard normal compliance point measurements in **Step 4** of the intrawell simulation, a set of such random observations needs to be generated for *each* of the *w* wells in the network. The behavior of *w* control charts must be simulated using a common set of background data and single control limit for each one.

Once a control limit meeting the target SWFPR has been established, a second Monte Carlo simulation is run to determine the statistical power of the control chart. Since effective power is defined as the ability to flag a single contaminated well-constituent pair, the basic steps are the same for either interwell or intrawell control charts. Repeat the following over a large number of simulations ($N_{sim}$).

1. Determine the background sample size ($n$). Generate $n$ standard normal measurements. From these, form baseline estimates by computing the sample mean ($\bar{x}_B$) and standard deviation ($s_B$).

2. Using the standardized control limit ($h$) chosen in the first Monte Carlo simulation, compute a non-standardized control limit using equation [20.4]. Then simulate the behavior of the control chart from sampling event 1 to sampling event $M$ by generating sets of normal N(Δ,1) compliance measurements for each event, where Δ varies from 1 to 5 by unit steps. Generate enough random measurements in each set to account for resamples potentially needed with a selected retesting strategy.

3. For each set of successively higher-valued compliance measurements, test the initial measurement associated with each sampling event against the non-standardized control limit. Also form the CUSUM for events 1 to $M$ using equations [20.2] and [20.3]. Compare the non-standardized CUSUM against the control limit.

4. If either the initial measurement or the CUSUM exceeds $h_c$, use the resample(s) for that sampling event to perform a retest (see below). If the retest confirms the initial exceedance, record a true detection for that particular mean-level Δ and simulation (out of $N_{sim}$).

5. After all $N_{sim}$ runs have been conducted, compute the observed power (1–β) associated with each true mean level (Δ) by dividing $N_{sim}$ into the number of observed detections. The simulated effective power curve for standardized control limit ($h$) is a plot of (1–β) versus Δ for Δ = 1 to 5.

If the standardized control limit identified during Monte Carlo simulation has effective power comparable to the appropriate ERPC (matching the site-specific sampling frequency to one of the three curves in **Chapter 6**: quarterly, semi-annual, or annual), $h$ can be used to form site-specific control limits. For interwell limits, compute the (upgradient) background mean and standard deviation for each monitoring constituent and use equation [20.4] to form the final, non-standardized control limits. For intrawell limits, use the same equation only with intrawell background at each well-constituent pair.

## 20.4.2 RETESTING IN CONTROL CHARTS

Control chart and prediction limit tests are only practical for most monitoring networks if retesting is part of the procedure, demonstrated both by Gibbons (1999) and Davis (1999). A key issue is to decide how control chart retesting should be conducted. Practical retesting strategies for prediction

limits on future observations are described in **Section 19.1**, including both 1-of-*m* (for *m* = 2, 3, 4) and modified California plans.

ASTM Standard D6312-98 (1999) recommends a 1-of-2 retesting strategy: whenever an exceedance of the control limit occurs on a given sampling event, the next quarterly sampling event is used as the resample. Furthermore, if the exceedance is not confirmed by the resample, the ASTM standard recommends that the initial exceedance be *replaced* in the CUSUM by the follow-up sampling event, thus implicitly assuming that the initial observation was an error.

Gibbons (1999) considers the performance of other retesting plans, including 1-of-2, 1-of-3, and the original Cal-3 plan (see **Section 19.1** and **Appendix B**). For each plan, resampling is triggered when the most recent observation either by itself exceeds or causes the CUSUM to exceed the limit. Then, each resample (if more than one) is compared against *h*. The initial exceedance measurement is removed from the CUSUM computation, replaced by the resample, and then re-compared to the control limit. A statistically significant increase [SSI] is declared only if the resample verifies the initial exceedance (or both resamples for a 1-of-3 plan).

Gibbon's study and ASTM Standard D6312-98 raises an important concern as to the most statistically powerful treatment of the CUSUM when an initial exceedance is *not* confirmed by retesting. A second concern addresses when resamples should be collected.

The Unified Guidance suggests two practical possibilities to address the first concern. The initial exceedance can be removed from the CUSUM altogether, re-setting the CUSUM to its value from the previous sampling event. As noted above, this is essentially assuming the first sampling event was in error. Another option is to replace the initial exceedance by the first resample which disconfirms the exceedance, and then re-compute the CUSUM with that resample.

In either strategy, the effects on statistical power and accuracy should be simulated when constructing site-specific control limits as in the procedure outlined above. Both the false positive rate and power depend on a faithful simulation of *all* aspects of the control chart testing procedure. This includes background sample size, the number of well-constituent pairs evaluated, the retesting strategy and how the CUSUM is adjusted for resampling.

The second issue concerns when resamples should be collected. The Unified Guidance does not recommend using the next scheduled sampling event as a resample. If the exceedance were due to a laboratory analytical error or calculation mistake, a more quickly retrieved resample can resolve the discrepancy without waiting until the next quarterly or semi-annual monitoring event.

Where multiple resamples are used (a 1-of-3 plan, for instance), one would have to wait two additional sampling rounds simply to collect the resamples. These in turn could not be plotted on the control chart as regular sampling events without intermingling the roles of resamples and non-resamples, thereby complicating the interpretation and assessment of control chart performance. The common guidance recommendation is to identify an intermediate period or periods for resampling between regularly scheduled evaluations for both control charts and prediction limits.

*This page intentionally left blank*